FREEDOM THROUGH IMPERFECTION:

EXPLOITING THE FLEXIBILITY OFFERED BY

REDUNDANCY IN SIGNAL PROCESSING

RACHEL WARD

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE PROGRAM IN

APPLIED AND COMPUTATIONAL MATHEMATICS

ADVISER: INGRID DAUBECHIES

Abstract

This thesis consists of four chapters. The first two chapters pertain to the design of stable quantization methods for analog to digital conversion, while the third and fourth chapters concern problems related to compressive sensing.

In the first chapter, we study the β-encoder and golden ratio encoder, and show that these quantization schemes are robust with respect to uncertainty in the multiplication element of their implementation, whereby $x \to \beta x$. In particular, we use a result from analytic number theory to show that the possibly unknown value of β can be reconstructed as the unique root of a power series with coefficients in $\{-1, 0, 1\}$ obtained from the quantization output of test input x and its negative, $-x$.

The focus of our attention in Chapter 2 is Sigma Delta ($\Sigma\Delta$) modulation, a course quantization method in analog to digital conversion that is widely used for its simplicity of implementation and robustness to component imperfections. A persistent problem in $\Sigma\Delta$ quantization of audio signals is the occurrence of undesirable *tones* arising from periodicities in the bit output; these tones are particularly intolerable in the space between successive audio tracks. As one of the contributions of this thesis, we show that the standard second order 1-bit $\Sigma\Delta$ scheme can be modified so as to eliminate such periodicities, and this modification can be achieved without sacrificing accuracy or introducing significant complexity.

The emerging area of compressed sensing guarantees that sufficiently sparse signals

can be recovered from significantly fewer measurements than their underlying dimension suggests, based on the observation that an N dimensional real-valued vector can be reconstructed efficiently to within a factor of its best k-term approximation by taking $m = k \log N$ measurements, or inner products. However, the quality of such a reconstruction is not assured if the underlying sparsity level of the signal is unknown. In Chapter 3, we show that sharp bounds on the errors between a signal and p estimates (corresponding to a different sparsity hypotheses, say) can be achieved by reserving only $O(\log p)$ measurements of the total m measurements for this purpose. The proposed technique is reminiscent of cross validation in statistics and learning theory, and theoretical justification in the context of compressed sensing is provided by the Johnson Lindenstrauss lemma.

In Chapter 4, we study the relation between nonconvex minimization problems arising in compressed sensing and those known as *free-discontinuity* problems, which are often encountered in the areas of image segmentation and edge detection. We adapt recent thresholding techniques in the area of sparse recovery to a class of nonconvex functionals that contains both compressed sensing and free-discontinuity-type problems. Along the way, we establish a connection between these two types of problems, using the notion of gamma convergence, and gain new insights into the minimizing solutions of such functionals.

Acknowledgements

First and foremost, I would like to thank my graduate advisor, Ingrid Daubechies, for directing me towards interesting and challenging problems, and for giving me the confidence to attack them. You are truly an inspiration, and I feel so proud and fortunate to have had the opportunity to work with you.

I would also like to thank the signal processing community as a whole for being friendly and warm to young people in the field. I have had many wonderful interactions and established nice collaborations as a result. In particular, I would like to thank Massimo Fornasier for being a great colleague and friend. Working with you is so much fun! Your attitude is infectious, and I can only hope to pass it along to others. I am also particularly thankful to Robert Calderbank, Sinan Gunturk, Albert Cohen, Thao Nguyen, and Ozgur Yilmaz for enlightening discussions that strengthened this thesis immensely.

I am thankful to the graduate students in PACM for making me feel part of a family. I have yet to encounter another department that gets along as well as we do. I am thankful to Arie for helping me simplify several of the arguments in this thesis, but more importantly for being by my side during the dark times, and for yelling at me when I take work too seriously. And finally, to my amazing family: your love and support inspires and strengthens me every day.

To my father, for introducing me to the strange and wonderful world of mathematics.

Contents

.

"But what did the bird see in the clear stream below? His own image; no longer a dark, gray bird, ugly and disagreeable to look at, but a graceful and beautiful swan."

- The Ugly Duckling

0.1 Overview

The first part of the thesis is concerned with the acquisition of real-valued, continuous-time signals, the likes of which we unknowingly process every day: speech, music, video, and wireless communications are all produced in this form, for example. At the same time, it is becoming increasingly efficient to store and manipulate information in *digital* format. This is partly due to the greater robustness of digital signals versus analog signals with respect to noise and fluctuation: any slight variation in an analog signal can cause a considerable amount of distortion, but for digital sequences can be safely rounded off to the nearest element in the discrete alphabet. The design and implementation of fast, stable, and accurate algorithms for conversion between the continuous and discrete domain is therefore crucial.

The process of *analog-to-digital conversion* usually consists of two parts: *sampling* and *quantization*. Sampling consists of converting the continuous-time signal of interest $f(t)$ into a discrete-time signal $f(t_n)$. According to the Shannon - Nyquist interpolation formula, this operation is invertible if the analog signal of interest is uniformly bounded and absolutely integrable, $f \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$, and *bandlimited*, meaning that its Fourier transform $\hat{f}(\omega) = \int_{-\infty}^\infty f(t) \exp{(-i\omega t)} dt$ vanishes outside a bounded interval $[-\Omega/2, \Omega/2]$. The importance of this result lies in the fact that many signals of practical engineering interest are well-approximated by bandlimited functions. For example, speech signals can be modelled as bandlimited functions whose bandwidth Ω is 4 KHz, while audio signals are well modelled as bandlimited functions whose bandwidth is 20 KHz [59].

Specifically, Ω-bandlimited functions $f \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$ can be recovered from sufficiently dense evenly-spaced samples according to the following reconstruction formula, which holds for $\lambda \geq 1$:

$$f(t) = \frac{1}{\lambda} \sum_n f\left(\frac{\pi n}{\lambda\Omega}\right) g\left(t - \frac{\pi n}{\lambda\Omega}\right); \tag{1}$$

here, the reconstruction filter g can be any function satisfying $\|\hat{g}\|_\infty \leq 1$, $\hat{g}(\omega) = 1$ for $|\omega| \leq \Omega/2$, and $\hat{g}(\omega) = 0$ for $|\omega| \geq \lambda\Omega/2$. At critical sampling rate $\lambda = 1$, or *Nyquist rate*, these restrictions determine the Fourier transform $\hat{g}$ as the indicator function for the interval $[-\Omega/2, \Omega/2]$, yielding the *sinc* filter, $g(t) = \Omega\text{sinc}(\Omega t) = \sin(\Omega\pi t)/\pi t$. As $\text{sinc}(t)$ is not absolutely summable, sampling at the critical rate $\lambda = 1$ renders the reconstruction (1.1) unstable in the presence of inevitable additive error on the samples; if noisy input $f_n = f(\frac{\pi n}{\Omega}) + \epsilon_n$ is instead observed, and $|\epsilon_n| = \epsilon$, then the sign pattern of the noise sequence (ϵ_n) can be chosen adversarily so that the resulting series $\sum_n f_n g(t - \frac{\pi n}{\Omega})$ does not even converge [31].

In practice, reconstruction is stabilized by oversampling at a fixed rate corresponding to $\lambda > 1$. In this setting, one has the freedom to design g so that its Fourier transform $\hat{g}$ is smooth, in which case g will have fast decay, and only those samples $f(\frac{\pi n}{\Omega})$ for which $|\frac{\pi n}{\Omega} - t|$ is sufficiently small will contribute significantly towards the reconstruction of $f(t)$.

Quantization in analog to digital conversion

The second step in analog to digital conversion is quantization, whereby the real-valued sequence $(f_n^\lambda) = \left(f(\frac{\pi n}{\lambda \Omega})\right)$ is encoded in a bitstream $(q_{(n,m)})$, with K bits $(q_{(n,m)})_{m=1}^K$ allocated to each sample f_n^λ. From these bits, the original function $f(t)$ can be reconstructed in the analog domain, albeit imperfectly, by replacing each sample f_n^λ in (1.1) by a function of the K-bit sequence $(q_{(n,m)})_{m=1}^K$. For example, this sequence could be taken to be the first K bits in the binary expansion of f_n^λ - such *pulse code modulation* quantization schemes are often employed in practice. In Chapter 1, we will study a related quantization scheme, the β-encoder, whereby expansions of f_n^λ in base $1 < \beta < 2$ are considered in place of binary expansions (where $\beta = 2$). Note that quantization, unlike sampling, is no longer an invertible operation. However, in any reasonable quantization scheme, the accuracy of the approximation to f from the quantization sequence $(q_{(n,m)})$ will increase as as function of $B = K\lambda$, the number of bits spent per Nyquist interval. That is, oversampling allows for more accurate as well as more stable reconstruction, but at the expense of additional sampling resources. The balance between accuracy, stability, and sampling efficiency in signal acquisition will be a recurring theme throughout this thesis.

In Chapters 1 and 2, we focus on the design of stable quantization schemes, assuming that we may spend many bits per Nyquist interval in acquiring samples (f_n^λ). This is the case for audio signal processing, where signals of interest have reasonable bandwidth $\Omega \leq 40$ KHz, and sampling at a rate of e.g. 32 bits per Nyquist interval corresponds to using fewer than 1 million bits per second. In determining the stability of a particular quantization scheme to nonidealities in its implemen-

tation, of particular importance is its sensitivity to imperfections in the *quantizer* element $Q : \mathbb{R} \to \{-1, 1\}$ that is necessary for conversion from the analog to digital representation. Q is usually taken to be the signum function $Q(u) = \text{sgn}(u)$, the behavior of which an actual quantizer, having finite precision and subject to thermal fluctuations, can only approximate. Actual quantizer elements usually come with a pre-assigned tolerance ϵ: for input with magnitude below this tolerance, $|u| \leq \epsilon$, the output $Q_\epsilon(u) \in \{-1, 1\}$ is not reliable. We will say that a quantization scheme is *robust* with respect to quantization error if, for some $\epsilon > 0$, the worst approximation error produced by the quantization scheme can be made arbitrarily small by allowing a sufficiently large bit budget, even if the quantizer used in its implementation is imprecise to within a tolerance ϵ. Such robustness can be achieved if the quantization sequence $(q_{(n,m)})$ constitutes a *redundant* representation of the original $f(t)$, so that an incorrect bit $q'_{(m,n)}$ caused by quantization error can be redeemed by an appropriate choice of subsequent bits. Of course, any quantization scheme of interest will require additional components in its implementation, such as adders, multipliers, integrators, and so forth; robustness with respect to imperfections in all these elements must be taken into account in evaluating the reconstruction accuracy of a particular scheme.

In Chapter 1, we investigate robustness properties of the β-encoder, a quantization scheme that was recently designed to be robust with respect to quantization errors, while also having asymptotically optimal reconstruction accuracy. We show that the β-encoder is also robust with respect to the other components in its implementation,

securing the validity of its reconstruction accuracy. Interestingly, such robustness is achieved by enforcing a slight shift in the quantizer element, $Q(u) = \operatorname{sgn} u + \tau$; that is, we take advantage of the *freedom to choose* an imperfect quantizer, offered by redundancy, in order to ensure additional robustness guarantees.

In Chapter 2, we focus on a different notion of stability in the design of quantization schemes. Sigma Delta ($\Sigma\Delta$) quantization, and in particular one-bit $\Sigma\Delta$ modulation, is a *coarse* quantization method, as each sample value f_n^λ in (1.1) is replaced by a single bit $q_n \in \{-1, 1\}$ that depends on all previous f_k^λ. Because of their robustness properties and ease of implementation, $\Sigma\Delta$ schemes are widely popular in practice, despite suffering from an entirely different kind of problem: in audio signal processing, periodic oscillatory patterns in the bit output q_n often occur, producing idle tone components in response to stretches of zero input $f_n^\lambda = 0$, or even small amplitude sinusoidal inputs. Such tones are audible to the listener, and are so pronounced in low-order $\Sigma\Delta$ quantizers that such schemes are often avoided in audio applications for more complex higher order $\Sigma\Delta$ schemes [51]. As one of the contributions of this thesis, we introduce a second-order $\Sigma\Delta$ scheme that eliminates such periodicities by damping the internal variables of the system once the input f_n^λ is identically zero. This approach is in line with the philosophy of Chapter 1: to eliminate the idle tones, we exploit the freedom to induce a small amount of leakage to the system without sacrificing accuracy, as offered by redundancy of the $\Sigma\Delta$ quantization output.

Beyond Nyquist sampling

The second part of the thesis concerns *compressed sensing*, a fast emerging area of applied mathematics that represents a new approach to traditional sampling paradigms. At the core of this field is the remarkable fact that a large class of underdetermined linear systems $y = \Phi x$ are invertible, and can be inverted efficiently, subject to the constraint that x is sufficiently *sparse*, or has sufficiently small nonzero support $\|x\|_0 = |\{i : x_i \neq 0|\}$. In particular, a matrix $\Phi \in \mathbb{R}^{m \times N}$ with unit normed columns is said to have the (k, δ)-*restricted isometry property*, or satisfy k-RIP for short, if all singular values of any k column submatrix of Φ lie in the interval $[1 - \delta, 1 + \delta]$ for a given constant $\delta < 1$. For $2k$-RIP matrices, a k-sparse vector x can be recovered from its lower-dimensional image $y = \Phi x$ as the minimal ℓ_1 norm solution,

$$x = \arg \min_{\Phi u = y} \|u\|_{\ell_1^N}. \tag{2}$$

That is, $2k$-RIP for the matrix Φ furnishes an equivalence between the minimal ℓ_0 norm and minimal ℓ_1 norm solution from the affine space $\{u \in \mathbb{R}^N : \Phi u = y\}$ if a k-sparse solution exists; while finding the former solution represents an NP-hard problem in general, the latter solution can be recovered efficiently using linear programming methods. Even more can be said: for vectors x satisfying $y = \Phi x$ that are approximately but not exactly k-sparse, the error $\|x - \hat{x}\|_{\ell_2^N}$ incurred by the approximation $\hat{x} = \arg\min_{\Phi u = y} \|u\|_1$ using $2k$-RIP matrix Φ is still small, thanks to the following stability result:

$$\|x - \hat{x}\|_{\ell_2^N} \leq \frac{C_\delta}{\sqrt{k}} \inf_{\|z\|_0 \leq k} \|x - z\|_{\ell_1^N}. \tag{3}$$

7

In words, the error $\|x - \hat{x}\|_{\ell_2^N}$ is bounded by a fraction of the best possible approximation error between x and the set of k-sparse vectors in the metric of ℓ_1^N [15].

Constructing k-RIP matrices With high probability, an $m \times N$ random matrix whose coefficients are drawn as independent and identical realizations of a Gaussian or Bernoulli random variable will satisfy k-RIP of optimal order,

$$k = O\big(m/\log\left(N/m\right))\big). \tag{4}$$

While no deterministic constructions of RIP matrices have been found of this order, many matrices with a smaller degree of randomness satisfy the restricted isometry property to almost optimal order. Most notably, with high probability an $m \times N$ matrix obtained by selecting m rows at random from the $N \times N$ discrete Fourier matrix satisfies k-RIP of order $k = O\big(m/(\log N \log\left(\log N\right))\big)$ [66]. This particular result has an interesting interpretation that connects back to the first two chapters: signals that are well-approximated as elements from the class of *sparse trigonometric polynomials*,

$$f(t) = \sum_{\omega=-N}^{\omega=N} a_\omega \exp\left(2\pi i \omega t\right), \; t \in [0,1) \; , \; a_\omega \in \mathbb{R}, \|a\|_0 \le d, \tag{5}$$

can be efficiently reconstructed from only $m = O(d \log^3 N))$ samples; when d is small, this represents an *exponentially smaller* number than the $O(N)$ samples required for the Shannon-Whittaker reconstruction formula (1.1) to hold. This is significant in applications such as radar, navigation, and satellite communications, where signals

of interest often have a sparse frequency representation, while measurements are at the same time expensive to implement.

Cross validation in compressed sensing

From the reconstruction formula (3) and order relation (4), the quality of the approximation $\hat{x} = \arg\min_{\Phi u = y} \|u\|_1$ to an unknown signal x in the affine space $\Phi x = y$ depends on the approximability of x is by its $k = \frac{m}{\log N/m}$ largest coordinates. In the literature, a known and fixed value for k is assumed to well-approximate all signals in the class of interest, while more realistically, k may be a parameter of the unknown input. Natural images, for instance, are generally well-approximated by only a few discrete Wavelet basis elements [37]; however, certain heavily textured images, such as those containing hair or sand, are not particularly compressible in such bases. In statistics, parameter selection and noise estimation are commonly achieved through *cross validation*, whereby the available data is separated into independent *testing* and *training* sets. In Chapter 3, we show that cross validation incorporates naturally into the compressed sensing paradigm, as the random measurements often used for compressed sensing are the same measurements that provide almost-isometric lower dimensional embeddings of generic point sets, as guaranteed by the Johnson-Lindenstrauss lemma. More precisely, we take a set of m measurements of the unknown x, and use $m - r$ of these measurements, Φx, in a compressed sensing decoding algorithm to return a sequence $(\hat{x}_1, \hat{x}_2, ...)$ of candidate approximations to x. The remaining r measurements, Ψx, are then used to identify from among this sequence a 'best' approximation $\hat{x} = \hat{x}_j$, along with an estimate of the sparsity

9

level of x. The proposed method for error estimation in compressed sensing is *extremely cheap*: approximation errors of up to p distinct approximations $(\hat{x}_j)_{j=1}^p$ can be estimated with high accuracy at the expense of only $10 \log p$ samples. Whereas in Chapters 1 and 2 we exploited redundancy of representation afforded by over-sampling to provide stability results for quantization schemes, we now exploit the stability of random measurements to validate the assumption that the representation at hand is redundant.

Compressed sensing and free-discontinuity problems

Chapter 4 explores the connection between minimization problems arising in compressed sensing and those corresponding to *free-discontinuity problems*, which describe situations where the solution of interest is defined by a function and a lower dimensional set consisting of the discontinuities of the function. Such situation arise, for instance, in crack detection from fracture mechanics [65] or in certain digital image segmentation problems [41]. The best-known example of a free-discontinuity problem is that of minimizing the so-called Mumford-Shah functional [22], which is defined by

$$J(u, K) := \int_{\Omega \setminus K} \left[|\nabla u|^2 + \alpha(u - g)^2 \right] dx + \beta \mathcal{H}^{d-1}(K \cap \Omega);$$

here, the set Ω is a bounded open subset of $\mathbb{R}^d$, the constants $\alpha, \beta > 0$ are fixed, $g \in L^\infty(\Omega)$, and $\mathcal{H}^N$ denotes the N-dimensional Hausdorff measure. In the context of visual analysis, the function g is a given noisy image that is to be approximated by the minimizing function $u \in W^{1,2}(\Omega \setminus K)$; the set K is simultaneously is used in

order to *segment* the image into connected components.

The Mumford-Shah functional $J(u, K)$ is not easily handled, theoretically or numerically speaking. This difficulty has given rise to the development of approximation methods for the Mumford-Shah functional and its minimizers where *sets* are no longer involved [22]. In the one-dimensional setting, minimizers of $J(u, K)$ can be approximated in a strong sense by minimizers of a discrete functional which, reformulated in terms of the discrete derivative, and generalized to incorporate inverse problems, takes the form of a *selective least squares* problem,

$$\text{Minimize } \mathcal{J}_r(u) := \|Tu - f\|_{\ell_2^m}^2 + \sum_{i=1}^{N} \min\left\{|u_i|^2, r\right\}. \tag{6}$$

Note that the unknown discrete derivative vector, assumed to be 'small' and smoothly varying except on the discontinuity set of the solution, should be well-approximated by the minimizer of $\mathcal{J}_r$. However, despite successful numerical results [24] observed for such selective least squares problems, no rigorous results on the existence of minimizers, let alone on the convergence of several proposed algorithms to such minimizers, are currently available in the literature.

In Chapter 4, we establish a connection between the selective least squares problem (6) and the ℓ_0 minimization problem from compressed sensing,

$$\text{Minimize } \mathcal{J}_r(u) := \|Tu - f\|_{\ell_2^m}^2 + \|u\|_0 \tag{7}$$

11

using the notion of gamma convergence. Moreover, we derive an iterative thresholding algorithm for finding solutions to the selective least squares functional $\mathcal{J}_r$, motivated by the recent application of such algorithms for minimizing the ℓ_0 regularized functional (7). Our algorithm is shown to converge to local minimizers $\bar{u}$ of the functional $\mathcal{J}_r$ that are characterized by certain fixed point conditions, including the following *gap* property: for each j, either $|\bar{u}_j| \geq \sqrt{2}r$ or $|\bar{u}_j| \leq \frac{1}{\sqrt{2}}r$. We show in addition that any global minimizer of $\mathcal{J}_r$ must satisfy such fixed point conditions, giving mathematical justification to the observation that minimizers of $\mathcal{J}_r$ tend to be "cartoon"-like, or segmented into regions of small gradient, separated by edges corresponding to large gradient. Moreover, we show that minimizers are restricted to regions where the non-convex functional $\mathcal{J}_r$ is locally strictly convex. Thus, global minimizers are always isolated, although not necessarily unique, whereas local minimizers may constitute a continuum of unstable equilibria. These observations sheds light on fundamental properties, virtues, and limitations, of regularization by means of the Mumford-Shah functional, and provide a rigorous justification of the numerical results appearing in the literature.

Contributions of this thesis

Chapter 1. The β-encoder and golden ratio encoder were recently introduced as quantization schemes in analog to digital conversion that are simultaneously robust with respect to quantizer imperfections in their analog implementation and efficient in their bit-budget use. However, it was not clear at first [34] that these schemes were also robust with respect to imprecisions in the multiplier element. Using a result in analytic number theory concerning the roots of power series with coefficients in $\{-1, 0, 1\}$, we show that these schemes are indeed robust as such, and the underlying value of β can be reconstructed as the unique root on $[0, 1]$ of a power series obtained from the quantization output of test input $(x, -x)$.

Chapter 2. Sigma Delta ($\Sigma\Delta$) modulation is a course quantization method in analog to digital conversion that is widely used for its simplicity of implementation and robustness properties; however, a persistent problem in $\Sigma\Delta$ modulation is the occurrence of undesirable idle tones arising from oscillatory patterns in the bit output. Such oscillations are omnipresent, and automatically arise along stretches of zero input, the likes of which are unavoidable in audio signal processing. Responding to a question left open in [79], we introduce a family of *quiet* second-order 2-bit asymmetrically damped $\Sigma\Delta$ schemes whose quantization output becomes constant at the onset of zero input, effectively eliminating the idle tones that arise in this context.

Chapter 3. The emerging area of compressed sensing boasts efficient sensing techniques based on the phenomenon that an N dimensional real-valued vector can be

reconstructed efficiently to within a factor of its best k-term approximation error by taking only $m = k \log N$ measurements. However, the quality of such a reconstruction is not assured if the underlying sparsity level of the signal is unknown. We show that sharp bounds on the errors between a signal and p approximations to the signal (corresponding to a different sparsity hypotheses, say) can be achieved by reserving only $O(\log p)$ measurements of the total m measurements for this purpose. This error estimation technique is reminiscent of cross validation in statistics and learning theory, and theoretical justification in the context of compressed sensing is provided by the Johnson Lindenstrauss lemma.

Chapter 4. Inspired by recent thresholding techniques in compressed sensing, we develop an algorithm for minimizing a discrete version of the Mumford Shah (MS) functional arising in image segmentation [21]. Although regularization methods involving similar nonconvex functionals work well in practice [63], ours is the first proven to converge to local minimizers of the discrete MS functional. The proposed algorithm can be adapted to a more general class of nonconvex functionals that includes the ℓ_0-regularized functional from compressed sensing. Finally, we show that the ℓ_0 regularized functional can be viewed as the limit of a sequence of free-discontinuity type problems, illuminating an intimate connection between the two problems.

Chapter 1

Robust quantization in analog to digital conversion

1.1 Analog to digital conversion: an introduction

Digital signals are omnipresent; one important reason is that transmission and storage for digital signals is much more robust with respect to noise and fluctuation than for analog signals. Because many signals of interest to us are produced in analog form, a transition from analog to digital is therefore necessary.

Analog-to-digital (A/D) conversion consists of two parts: *sampling* and *quantization*. The process of sampling, followed by quantization, can be schematically represented by

$$f(t) \Rightarrow \left(f(t_n)\right)_{n \in Z} \Rightarrow \left(\left(q_{(n,m)}\right)_{m=1}^{K}\right)_{n \in Z}$$

where $q_{(n,m)} = \left(q_{(n,1)}, q_{(n,2)}, ..., q_{(n,K)}\right)$ is a finite sequence of digits from a finite (and usually binary) alphabet. It is reasonable to assume that the real-valued analog signal of interest, f, is uniformly bounded and absolutely integrable, $f \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$, and is *bandlimited*; i.e., its Fourier transform $\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t) \exp(-i\omega t) dt$ vanishes outside some bounded interval $[-\Omega/2, \Omega/2]$. Under this assumption, the sampling step in A/D conversion is an invertible operation, as $f(t)$ can be reconstructed from the samples $\left(f(\frac{n\pi}{\Omega\lambda})\right)_{n \in Z}$ as long as $\lambda \geq 1$, via the reconstruction formula

$$f(t) = \frac{1}{\lambda} \sum_n f\left(\frac{n\pi}{\lambda\Omega}\right) g\left(t - \frac{n\pi}{\lambda\Omega}\right). \tag{1.1}$$

Above, g can be any function such that $\|g\|_\infty \leq 1, \hat{g} = 1$ for $|\xi| \leq \Omega$ and $\hat{g} = 0$ for $|\xi| \geq \lambda\Omega$. If $\hat{g}$ is smooth, then g will have fast decay and the reconstruction (1.1) will be almost entirely local [31]. By scaling and normalizing appropriately, we can reduce any bandlimited function in $L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$ to an element of the class $S = \{f : \|f\|_\infty \leq 1, \|f\|_1 \leq 1, \text{supp } \hat{f} \in [-\pi, \pi]\}$, in which case the reconstruction formula (1.1) reduces to

$$f(t) = \frac{1}{\lambda} \sum_n f\left(\frac{n}{\lambda}\right) g\left(t - \frac{n}{\lambda}\right). \tag{1.2}$$

The second step in A/D conversion is *quantization*, in which sample values $f(\frac{n}{\lambda})$ are replaced by finite bitstreams $(q_{(n,m)})_{m=1}^{K}$ and associated function $\Delta : \{0, 1\}^K \to \mathbb{R}$, so that the original function can be reconstructed at a later time by the approximation

$$\tilde{f}(t) = \frac{1}{\lambda} \sum_n \Delta(q_{(n,.)}) g\left(t - \frac{n\pi}{\lambda\Omega}\right). \tag{1.3}$$

Quantization, unlike sampling, is not in general an invertible operation. That is, if E is the operator which maps functions in S to bitstreams (the *encoding* operator) associated with a particular quantization scheme, and D is the *decoding* operator which maps bitstreams back to functions in S, then typically $f \neq D(Ef)$. In order to measure the performance of a particular quantization scheme, one considers the *distortion* of the scheme, given by $d(S; E, D) := \sup_{f \in S} \|f - D(Ef)\|$, where $\|.\|$ is the norm of interest; e.g.,the L^∞ norm on $\mathbb{R}$. In any reasonable quantization scheme, the distortion will decrease as the number of bits per unit interval (the so-called *bit budget*) increases; schemes whose distortion decreases faster as a function of the bit budget are generally considered superior quantization schemes.

If the relation between the distortion d and the bit budget B were the only important measure associated with a quantization scheme, then the widely-used *pulse code modulation* (PCM) quantization algorithm would always be preferred in practice. Given a function f in S, a N-bit PCM algorithm simply replaces each sample value $f(\frac{n}{\lambda})$ with N bits: one bit for its sign, followed by the first $N-1$ bits of the binary expansion of $|f(\frac{n}{\lambda})|$. One can show that for signals f in S, and for fixed $\lambda > 1$, the distortion $d(f, E, D) \leq C_\lambda 2^{-N} = C_\lambda 2^{-B/\lambda}$ for an N-bit PCM, when, for instance, $\|.\|$ is the L^∞ norm on $\mathbb{R}$.

Sigma Delta ($\Sigma\Delta$) quantization

Another popular quantization algorithm, $\Sigma\Delta$ *modulation*, has distortion that decays only like an inverse polynomial in the bit budget B. One-bit $\Sigma\Delta$ modulation replaces each sample f_n^λ in (1.1) by a single bit, $b_n \in \{-1, 1\}$ (In multibit $\Sigma\Delta$ schemes, the

17

coefficients $b(n)$ replacing the f_n^λ can assume a larger range of discrete values, and thus require several bits). In contrast to PCM, the b_n in $\Sigma\Delta$ depend not only on f_n^λ, but on all previous f_k^λ, $k \leq n$. Consider first-order, one-bit $\Sigma\Delta$, where the b_n are generated recursively by the scheme,

$$
\begin{aligned}
b_n &= Q(u_{n-1} + f_n^\lambda) \\
u_n &= u_{n-1} + f_n^\lambda - b_n.
\end{aligned}
\tag{1.4}
$$

The *quantizer* $Q(x) = \text{sgn}(x)$ just returns the sign of its argument, which can be taken to be either 1 or -1 when $x = 0$. Assuming $|f_n^\lambda| \leq 1$ (which is true of functions $f \in S$), and initializing $|u_0| < 1$, a simple inductive argument guarantees that $|u_n| < 1$ for all n [31]. Using this, along with the observation that $B = \lambda$ for one-bit quantization schemes, we arrive at an error estimate $\|f - \tilde{f}^B\|_\infty \leq 2\|\frac{dg}{dt}\|_1 B^{-1}$ for the first-order $\Sigma\Delta$ scheme (1.4). More generally, the Kth order analog of the recursion (1.4) equates $f_n^\lambda - b_n$ to the Kth order (as opposed to first order) difference of a bounded sequence $(u_n)_{n\in\mathbb{Z}}$, and has corresponding decay $\|f - \tilde{f}^B\|_\infty \leq C_{K,g} B^{-K}$; in particular, the second-order one-bit $\Sigma\Delta$ quantization scheme can be recast in the form

$$
\begin{aligned}
b_n &= Q(F(u_{n-1}, v_{n-1}, f_n^\lambda)) \\
u_n &= u_{n-1} + f_n^\lambda - b_n \\
v_n &= v_{n-1} + u_n.
\end{aligned}
\tag{1.5}
$$

The function F can be any map that guarantees the boundedness of the state variables (u_n, v_n); for example, the linear function $F(u_n, v_n) = u_n + \gamma v_n$ suffices for an appropriate range of γ [79].

Robustness for $\Sigma\Delta$ quantization

Clearly, the $O(2^{-B})$ decay of PCM will outperform the $O(B^{-K})$ decay of a K-th order one-bit $\Sigma\Delta$ scheme, for any order K. Yet, $\Sigma\Delta$ schemes of as low as second order are preferred in practice over high order PCM quantization schemes. This can be partly understood by the fact that $\Sigma\Delta$ schemes are less sensitive than PCM schemes to inevitable errors in the analog components of their implementation. The notion of *robustness* of $\Sigma\Delta$ schemes to quantization error, which was first studied in [31], is connected to how well $\Sigma\Delta$ schemes exploit the redundancy of the representation $f(t) = \frac{1}{\lambda}\sum_n c_n g(t - t_n^\lambda)$ afforded by oversampling $\lambda > 1$. By redundancy, we refer to the fact that the functions $\left(g_n(t)\right)_{n\in\mathbb{Z}} = \left(g(t - t_n^\lambda)\right)_{n\in\mathbb{Z}}$ form a frame [1] for the Hilbert space of Ω-bandlimited functions, and their frame redundancy increases with λ. One source of quantization error incurred in both $\Sigma\Delta$ and PCM quantization comes from the quantizer function itself: the function $Q(u) = \operatorname{sgn} u$ that is used in both the $\Sigma\Delta$ schemes and in the recursive algorithms used to generate binary expansions in PCM cannot be built to have infinite precision. In fact, quantizer elements for A/D circuits generally come with a prescribed tolerance ν for which the output $Q(u)$ of such quantizers should not be trusted once $|u| < \nu$ (and of course, quantizers with

---

[1]Technically, the shifts of g are not a frame for the space of Ω-bandlimited functions because these functions don't live in this space. But this is only a technicality due to the definition of a frame; the function g lives in the larger Hilbert space of $\Omega\lambda$-bandlimited functions, or simply L^2 for that matter, and its shifts satisfy the frame property for the smaller space of Ω-bandlimited functions.

lower tolerance are more expensive). As the binary expansion of almost every real number is unique, an incorrect bit assignment, and especially an incorrect initial bit, in the truncated binary expansion of a sample f_n^λ will cause an error in the resulting approximation that obviates the possibility of $O(2^{-B})$ decay. $\Sigma\Delta$ schemes, on the other hand, keep track of prior samples f_k^λ, $k \leq n$, in such a way that errors caused by imprecise quantizers can be corrected later, and the $O(B^{-K})$ maintained, by exploiting the redundant information in the samples f_n^λ [31].

This discussion brings us to the question: *Is it possible to have the best of both worlds*? That is, can one design a quantization scheme that has exponential reconstruction guarantees of PCM while also being robust with respect to quantization imperfections like $\Sigma\Delta$? This question was answered affirmatively in [33], with the introduction of the β-encoder.

1.2   On the robustness of beta-encoders and golden ratio encoders

Beta-encoders are similar to PCM in that they replace each sampled function value $x = f_n^\lambda$ with a truncated series expansion in a base β, where $1 < \beta < 2$, and with binary coefficients. Clearly, if $\beta = 2$, then this algorithm coincides with PCM. However, whereas the binary expansion of almost every real number is unique, for every $\beta \in (1, 2)$, there exist a continuum of different β expansions of almost every x in $(0, 1]$ (see [69]). It is precisely this redundancy that gives beta-encoders the freedom

20

to correct errors caused by imprecise quantizers shared by $\Sigma\Delta$ schemes. Whereas in $\Sigma\Delta$, a higher degree of robustness is achieved via finer sampling, beta-encoders are made more robust by choosing a smaller value of β as the base for expansion.

Although beta-encoders as discussed in [33] are robust with respect to quantizer imperfections, these encoders are not as robust with respect to imprecisions in other components of their circuit implementation. Typically, beta-encoders require a multiplier in which real numbers are multiplied by β. Like all analog circuit components, this multiplier will be imprecise; that is, although a known value β_0 may be set in the circuit implementation of the encoder, thermal fluctuations and other physical limitations will have the effect of changing the true multiplier to an unknown value $\beta \in [\beta_{low}, \beta_{high}]$ within an interval of the pre-set value β_0. The true value β will vary from device to device, and will also change slowly in time within a single device. This variability, left unaccounted for, disqualifies the beta-encoder as a viable quantization method since the value of β must be known with exponential precision in order to reconstruct a good approximation to the original signal from the recovered bit streams.

We overcome this potential limitation of the beta-encoder by introducing a method for recovering β from the encoded bitstreams of a real number $x \in [-1, 1]$ and its negative, $-x$. Our method incorporates the techniques used in [34], but our analysis is simplified using a transversality condition, as defined in [70], for power series with coefficients in $\{-1, 0, 1\}$. As the value of β can fluctuate within an interval $[\beta_{low}, \beta_{high}]$ over time, our recovery technique can be repeated at regular intervals

during quantization (e.g., after the quantization of every 10 samples).

The golden ratio encoder (GRE) was proposed in [35] as a quantizer that shares the same robustness and exponential rate-distortion properties as beta-encoders, but that does not require an explicit multiplier in its circuit implementation. GRE functions like a beta-encoder in that it produces beta-expansions of real numbers; however, in GRE, β is fixed at the value of the golden ratio, $\beta = \phi = \frac{1+\sqrt{5}}{2}$. The relation $\phi^2 = \phi + 1$ characterizing the golden ratio permits elimination of the multiplier from the encoding algorithm. Even though GRE does not require a precise multiplier, component imperfections such as integrator leakage in the implementation of GRE may still cause the true value of β to be slightly larger than ϕ; in practice it is reasonable to assume $\beta \in [\phi, 1.1\phi]$. Our method for recovering β in general beta-encoders can be easily extended to recovering β in the golden ratio encoder.

Our work in this section will be organized as follows:

1. In sections 1.2.1 and 1.2.2, we review relevant background on beta-encoders and golden ratio encoders, respectively.

2. In section 1.2.3, we introduce a more realistic model of the golden ratio encoder that takes into account the effects of integrator leak in the delay elements of the circuit. We show that the output of this revised model still correspond to truncated beta-expansions of the input, but in an unknown base β that differs from the pre-set value.

3. Section 1.2.4 describes a way to recover the unknown value of β up to arbitrary precision using the bit streams of a 'test' number $x \in [-1, 1]$, and $-x$. The recovery scheme reduces to finding the root of a polynomial with coefficients in $\{-1, 0, 1\}$.

4. Section 1.2.5 extends the recovery procedure of the previous section to the setting of beta-encoders having leakage in the (single) delay element of their implementation. We show that the analysis in this case is completely analogous to that of Section 1.2.4.

1.2.1 The beta-encoder

In this section, we summarize certain properties of beta-encoders (or β-encoders) with error correction, from the perspective of encoders which produce beta expansions with coefficients in $\{-1, 1\}$. For more details on beta-encoders, we refer the reader to [33].

We start from the observation that given $\beta \in (1, 2]$, every real number $x \in [-\frac{1}{\beta-1}, \frac{1}{\beta-1}]$ admits a sequence $(b_j)_{j \in N}$, with $b_j \in \{-1, 1\}$, such that

$$x = \sum_{j=1}^{\infty} b_j \beta^{-j}. \tag{1.6}$$

Under the transformation $\tilde{b}_j = \frac{b_j+1}{2}$, (1.6) is equivalent to the observation that every real number $y \in [0, \frac{1}{\beta-1}]$ admits a beta-expansion in base $\beta \in (1, 2]$, with coefficients $\tilde{b}_j \in \{0, 1\}$. Accordingly, all of the results that follow in this section have straight-

forward analogs in terms of $\{0,1\}$-beta-expansions; see [33] for more details.

One way to compute a sequence $(b_j)_{j \in N}$ that satisfies (1.6) is to run the iteration

$$
\begin{aligned}
u_1 &= \beta x \\
b_1 &= Q(u_1) \\
\text{for } j \geq 1 : u_{j+1} &= \beta(u_j - b_j) \\
b_{j+1} &= Q(u_{j+1})
\end{aligned}
\tag{1.7}
$$

where the quantizer Q is simply the sign-function,

$$
Q(u) = \begin{cases} -1, & u \leq 0 \\ 1, & u > 0. \end{cases}
\tag{1.8}
$$

For $\beta = 2$, the expansion (1.6) is unique for almost every $x \in [-1, 1]$; however, for $\beta \in (1, 2)$, there exist uncountably many expansions of the type (1.6) for any $x \in [-1, 1]$ (see [69]). Because of this redundancy in representation, beta encoders are robust with respect to quantization error, while PCM schemes are not. We now explain in more detail what we mean by quantization error. The quantizer Q in (1.8) is an idealized quantizer; in practice, one has to deal with quantizers that only approximate this ideal behavior. A more realistic model is obtained by replacing Q

24

in (1.8) with a 'flaky' version Q^ν, for which we know only that

$$
Q^\nu(u) = \begin{cases}
-1, & u < -\nu \\
1, & u \geq \nu \\
-1 \text{ or } 1, & -\nu \leq u \leq \nu.
\end{cases}
\tag{1.9}
$$

In practice, ν is a quantity that is not known exactly, but over the magnitude of which we have some control, e.g. $|\nu| \leq \epsilon$ for a known ϵ. This value ϵ is called the *tolerance* of the quantizer. We shall call a quantization scheme *robust* with respect to quantization error if, for some $\epsilon > 0$, the worst approximation error produced by the quantization scheme can be made arbitrarily small by allowing a sufficiently large bit budget, even if the quantizer used in its implementation is imprecise to within a tolerance ϵ. According to this definition, the naive $\{-1, 1\}$-binary expansion is not robust. More specifically, suppose that a flaky quantizer Q^ν is used in (1.7) to compute the base-2 expansion of a number $x \in [-1, 1]$ which is sufficiently small that $|2x| \leq \nu$. Since $2x$ is within the flaky zone for Q^ν, if $b_1 = Q^\nu(2x)$ is assigned incorrectly; i.e., if b_1 differs from the sign of x, then no matter how the remaining bits are assigned, the difference between x and the number represented by the computed bits will be at least $|x|$. This is not the case if $1 < \beta < 2$, as shown by the following whose proof can be found in [33]:

Theorem 1.2.1. *Let $\epsilon > 0$ and $x \in [-1, 1]$. Suppose that in the beta-encoding of x, the procedure (1.7) is followed, but the quantizer Q^ν is used instead of the ideal Q at each occurence, with ν satisfying $\nu \leq \epsilon$. Denote by $(b_j)_{j \in N}$ the bit sequence produced by applying this encoding to the number x. If β satisfies*

25

$$1 < \beta < \tfrac{2+\epsilon}{\epsilon+1},$$

then

$$|x - \sum_{j=1}^{N} b_j \beta^{-j}| \leq C\beta^{-N} \tag{1.10}$$

with $C = \epsilon + 1$.

For a given tolerance $\epsilon > 0$, running the recursion (1.7) with a quantizer Q^ν of tolerance ϵ and a value of β in $(1, \tfrac{2+\epsilon}{\epsilon+1})$ produces bitstreams (b_j) corresponding to a beta-expansion of the input x in base β; however, the precise value of β must be known in order recover x from such a beta-expansion. As mentioned in the introduction and detailed in the following section, component imperfections may cause the circuit to behave as if a different value of β is used, and this value will possibly be changing slowly over time within a known range, $[\beta_{low}, \beta_{high}]$. In [34], a method is proposed whereby an exponentially precise approximation $\tilde{\gamma}$ to the value of $\gamma = \beta^{-1}$ at any given time can be encoded and transmitted to the decoder without actually physically measuring its value, via the encoded bitstreams of a real number $x \in [0, 1)$ and $1 - x$. This decoding method can be repeated at regular time intervals during the quantization procedure, to account for the possible time-variance of γ. That an exponentially precise approximation $\tilde{\gamma}$ to γ is sufficient to reconstruct subsequent samples $f(t_n)$ with exponential precision is the content of the following theorem, which is essentially a restatement of Theorem 5 in [34].

Theorem 1.2.2 (Daubechies, Yilmaz). *Consider $x \in [0, 1)$ and $(b_j)_{j \in N} \in \{0, 1\}$, or $x \in [-1, 1]$ and $(b_j)_{j \in N} \in \{-1, 1\}$. Suppose $\gamma \in (1/2, 1)$ is such that $x = \sum_{j=1}^{\infty} b_j \gamma^j$. Suppose $\tilde{\gamma}$ is such that $|\gamma - \tilde{\gamma}| \leq C_1 \gamma^N$ for some fixed $C_1 > 0$. Then $\tilde{x}_N := \sum_{j=1}^{N} b_j \tilde{\gamma}^j$*

satisfies

$$|x - \tilde{x}_N| \leq C_2 \gamma^N \tag{1.11}$$

where C_2 is a constant which depends only on γ and C_1.

Although the approach proposed in [34] for estimating β from bitstreams overcomes potential approximation error caused by imprecise multipliers in the circuit implementation of the beta-encoder, new robustness issues are nevertheless introduced. Typically, one cannot ensure that the reference level 1 in $1-x$ is known with high precision. To circumvent this problem, the authors consider other heuristics whereby β is recovered from clever averaging of multiple pairs x_j and $1-x_j$. These heuristics do not require that the reference level 1 in $1-x$ be precise; however, these approaches become quite complicated in and of themselves, and any sort of analytical analysis of their performance becomes quite difficult. As one of the contributions of the present work, we present an approach for recovering β that is inspired by the approach in [34] but does not require a precise reference level, yet still allows for exponentially precise approximations to β.

1.2.2 The golden ratio encoder (GRE)

As shown in the previous section, beta-encoders are robust with respect to imperfect quantizer elements, and their approximation error decays exponentially with the number of bits N. To attain this exponential precision, β must be measured with high precision, which is quite complicated in practice. These complications motivated the invention of the golden ratio encoder (GRE) of [35], which has the same robustness

and rate-distortion properties as beta-encoders, but uses an additional delay element in place of precise multiplier in its implementation. That is, if one implements the recursion (1.7) with $\beta = \phi = \frac{1+\sqrt{5}}{2}$, then using the relation $\phi^2 = \phi + 1$, one obtains the recursion formula $u_{n+1} = u_{n-1} + u_n - (b_{n-1} + \phi b_n)$. If the term $b_{n-1} + \phi b_n$ in this formula is removed, then the resulting recursion $v_{n+1} = v_n + v_{n-1}$ should look familiar; indeed, with initial conditions $(v_0, v_1) = (0, 1)$, this recursion generates the Fibonacci numbers v_n, and it is well-known that the sequence $\frac{v_{n+1}}{v_n} \to \phi$ as $n \to \infty$. If $b_{n-1} + \phi b_n$ is instead replaced by a single bit taking values in $\{-1, 1\}$, then we are led to the following scheme:

$$
\begin{aligned}
u_0 &= x \\
u_1 &= 0 \\
\text{for } n \geq 0 : b_n &= Q(u_n, u_{n+1}) \\
u_{n+2} &= u_{n+1} + u_n - b_n
\end{aligned}
\tag{1.12}
$$

In this paper, we will consider quantizers Q in (1.12) of the form Q_α, where

$$
Q_\alpha(u, v) = \begin{cases} -1, & u + \alpha v < 0 \\ 1, & u + \alpha v \geq 0 \end{cases}
\tag{1.13}
$$

along with their flaky analogs,

$$Q_\alpha^\nu(u, v) = \begin{cases} -1, & u + \alpha v < -\nu \\ 1, & u + \alpha v \geq \nu \\ -1 \text{ or } 1, & -\nu \leq u + \alpha v \leq \nu \end{cases} \tag{1.14}$$

In [35], the authors consider the recursion formula (1.12) implemented with flaky $\{0, 1\}$-quantizers of the form $\bar{Q}_\alpha^{\nu,\iota}(u, v) = \left[Q^\nu(u + \alpha v - \iota) + 1 \right]/2$. For the simplicity of presentation, we will consider only the $\{-1, 1\}$-quantizers (1.14), but many of our results extend straightforwardly to quantizers of the type $\bar{Q}_\alpha^{\nu,\iota}$.

The following theorem was proved in [35]; it shows that as long as x and Q are such that the state sequence $u = \{u_n\}_{n=0}^\infty$ remains bounded, a golden ratio encoder (corresponding to the recursion (1.12)) will produce a bitstream (b_j) corresponding to a beta-expansion of x in base $\beta = \phi$, just as does the beta-encoder from which the GRE was derived.

Theorem 1.2.3 (Daubechies, Güntürk, Wang, Yilmaz). *Consider the recursion (1.12). Suppose the 1-bit quantizer Q which outputs bits (b_j) in $\{-1, 1\}$ is of the type Q_α^ν such that the state sequence $u = \{u_n\}_{n=0}^\infty$ with $u_0 = x$ and $u_1 = 0$ is bounded. Then*

$$\left| x - \sum_{n=0}^{N} b_n \phi^{-n} \right| \leq \phi^{-N+1} \tag{1.15}$$

Here ϕ is the golden ratio.

29

Proof. Note that

$$
\begin{aligned}
\sum_{n=0}^{N} b_n \phi^{-n} &= \sum_{n=0}^{N} (u_n + u_{n+1} - u_{n+2}) \phi^{-n} \\
&= \sum_{n=0}^{N} u_n \phi^{-n} + \sum_{n=1}^{N+1} u_n \phi^{-(n-1)} - \sum_{n=2}^{N+2} u_n \phi^{-(n-2)} \\
&= u_0 + (1 + \phi - \phi^2) \sum_{n=2}^{N} \phi^{-n} u_n \\
&\quad + u_1(\phi^{-1} + 1) + \phi^{-N}\left(u_{N+1}(1 - \phi) - u_{N+2} \right) \\
&= u_0 + \phi^{-N}\left(u_{N+1}(1 - \phi) - u_{N+2} \right) \\
&= x + \phi^{-N}\left(u_{N+1}(1 - \phi) - u_{N+2} \right).
\end{aligned}
$$

The second to last equality uses the relation $1 + \phi - \phi^2 = 0$, and the last equality is obtained by setting $u_0 = x$ and $u_1 = 0$. Since the state sequence $u = \{u_n\}_{n=0}^{\infty}$ is bounded, it follows that $x = \sum_{n=0}^{\infty} b_n \phi^{-n}$. Thus,

$$
\begin{aligned}
|x - \sum_{n=0}^{N} b_n \phi^{-n}| &= |\sum_{n=N+1}^{\infty} b_n \phi^{-n}| \\
&\leq \frac{\phi^{-(N+1)}}{1 - \phi^{-1}} \\
&= \phi^{-N+1}.
\end{aligned}
$$

$\square$

Although the implementation of GRE requires 2 more adders and one more delay element than the implementation of the beta-encoder, the multiplier element α in

GRE does not have to be precise (see section 6), whereas imprecisions in the multiplier element β of the beta-encoder result in beta-expansions of the input x in a different base β'.

1.2.3 GRE: A revised scheme incorporating integrator leak

In modeling the golden ratio encoder by the system (1.12), we are assuming that the delay elements used in its circuit implementation are ideal. A more realistic model would take into account the effect of integrator leak, which is inevitable in any practical circuit implementation (see [48] for more details). After one clock time, the stored input in the first delay is reduced to λ_1 times its original value, while the stored input in the second delay is replaced by λ_2 times its original value. In virtually all circuits of interest, no more than 10 percent of the stored input is leaked at each timestep; that is, we can safely assume that λ_1 and λ_2 are parameters in the interval [.9,1]. The precise values of these parameters may change in time; however, as virtually all practical A/D converters produce over 1000 bits per second (and some can produce over 1 billion bits per second), we may safely assume that λ_1 and λ_2 are constant throughout the quantization of at least every 10 samples.

Fixing an input value $x \in [-1, 1]$, we arrive at the following revised description

of the golden ratio encoder (revised GRE):

$$
\begin{aligned}
u_0 &= 0 \\
u_1 &= x \\
\text{for } n \geq 0 : b_n &= Q(\lambda_1\lambda_2 u_n, \lambda_1 u_{n+1}) \\
u_{n+2} &= \lambda_1\lambda_2 u_n + \lambda_1 u_{n+1} - b_n
\end{aligned}
\tag{1.16}
$$

Obviously, $\lambda_1 = \lambda_2 = 1$ corresponds to the original model (1.12). It is reasonable to assume in practice that $(\lambda_1, \lambda_2) \in M := [.95, 1]^2$, and in virtually all cases $(\lambda_1, \lambda_2) \in V := [.9, 1]^2$.

We will show that the revised scheme (1.16) still produces beta-expansions of the input x, but in a slightly different base $\gamma = \beta^{-1} = \frac{-\lambda_1 + \sqrt{\lambda_1^2 + 4\lambda_1\lambda_2}}{2\lambda_1\lambda_2}$, which increases away from ϕ^{-1} as the parameters λ_1 and λ_2 decrease. Key in the proof of Theorem 1.15 was the use of the relation $\phi^2 - \phi - 1 = 0$ to reduce $\sum_{n=0}^{N}(u_n + u_{n+1} - u_{n+2})\phi^{-n}$ to the sum of the input x, and a remainder term that becomes arbitrarily small with increasing N. Accordingly, the relation $1 - \lambda_1\gamma - \lambda_1\lambda_2\gamma^2 = 0$ gives $\sum_{n=0}^{N}(\lambda_1\lambda_2 u_n + \lambda_1 u_{n+1} - u_{n+2})\gamma^{n+1} = x + R(N)$, where $R(N)$ goes to 0 as N goes to infinity.

Theorem 1.2.4. *Suppose the 1-bit quantizer Q in (1.16) of type (1.14) is such that the state sequence $u = \{u_n\}_{n=0}^{\infty}$ with $u_0 = 0$ and $u_1 = x$ is bounded. Consider $\gamma = \frac{-\lambda_1 + \sqrt{\lambda_1^2 + 4\lambda_1\lambda_2}}{2\lambda_1\lambda_2}$. Then*

$$
|x - \sum_{n=0}^{N} b_n \gamma^{n+1}| \leq C_\gamma \gamma^N
$$

where $C_\gamma = \frac{\gamma}{1-\gamma}$.

32

Proof.

$$\sum_{n=0}^{N} b_n \gamma^{n+1} = \sum_{n=0}^{N}(\lambda_1 \lambda_2 u_n + \lambda_1 u_{n+1} - u_{n+2})\gamma^{n+1}$$

$$= (\lambda_1 \lambda_2 \gamma^2 + \lambda_1 \gamma - 1)\sum_{n=2}^{N-1} \gamma^n u_n$$

$$+ \lambda_1 \lambda_2 u_0 \gamma + u_1(\lambda_1 \gamma + \lambda_1 \lambda_2 \gamma^2)$$

$$+ \gamma^N \left(u_N(\lambda_1 - \gamma^{-1}) - u_{N+1} \right)$$

$$= x + \gamma^N \left(u_N(\lambda_1 - \gamma^{-1}) - u_{N+1} \right).$$

The last equality is obtained by setting $u_0 = 0$ and $u_1 = x$. Since the u_n are bounded, it follows as in the proof of (1.12) that

$$\left| x - \gamma \sum_{n=0}^{N} b_n \gamma^n \right| \leq \frac{\gamma^{N+1}}{1 - \gamma}.$$

$\square$

Theorem 1.2.4 implies that if γ is known, then the revised GRE scheme (1.16) still gives exponential approximations to the input signal x, provided that the u_n are indeed bounded. The following theorem gives an explicit range for the parameters (ν, α) which results in bounded sequences u_n when the input $x \in [-1, 1]$, independent of the values of the leakage parameters (λ_1, λ_2) in the set $V = [.9, 1]^2$. This parameter range is only slightly more restrictive than that derived in [35] for the ideal GRE scheme (1.12); that is, the admissable parameter range for α and ν is essentially robust with respect to leakage in the delay elements of the GRE circuit

33

implementation.

Theorem 1.2.5. *Let $x \in [-1, 1]$, and $(\lambda_1, \lambda_2) \in [.9, 1]^2$. Suppose that the GRE scheme* (1.16) *is followed, and the quantizer $Q_\alpha^\nu(u, v)$ is used, with ν possibly varying at each occurrence, but always satisfying $\nu \leq \epsilon$ for some fixed tolerance $\epsilon \leq .337$. If the parameter α takes values in the interval $[1.198 + 1.479\epsilon, 2.053 - 1.058\epsilon]$, then the resulting state sequence $(u_j)_{j \in N}$ is bounded.*

We leave the proof of Theorem 1.2.5 to the appendix. This result in some sense parallels Theorem 1.2.1 in that an admissable range for the "multiplication" parameter (α in GRE, β in beta-encoders) is specified for a given quantizer tolerance ϵ; however, we stress that the specific value of β is needed in order to recover the input from the bitstream (b_j) in beta-encoders. In contrast, a GRE encoder can be built with a multiplier α set at any value within the range $[1.198 + 1.479\epsilon, 2.053 - 1.058\epsilon]$, and as long as this multiplier element has enough precision that the true value of α will not stray from this interval, then the resulting bitstreams (b_j) will always represent a series expansion of the input x in base γ of Theorem 1.2.4, which does not depend on α. The base γ does however depend on the leakage parameters λ_1 and λ_2, which are not known a priori to the decoder and can also vary in time from input to input: as discussed earlier, the only information available to the decoder a priori is that $(\lambda_1, \lambda_2) \in [.9, 1]^2$ in virtually all cases of interest, and $(\lambda_1, \lambda_2) \in [.95, 1]^2$, or $\gamma \in [\phi^{-1}, \frac{20}{19}\phi^{-1}] \approx [.618, .6505]$ in most cases of interest. In the next section, we show that the upper bound of .6505 is sufficiently small that the value of γ can be recovered with exponential precision from the encoded bitstreams of a pair of real numbers $(x, -x)$. In this sense, GRE encoders are robust with respect to leakage in

34

the delay elements, imprecisions in the multiplier α, *and* quantization error.

1.2.4 Determining $\gamma = 1/\beta$ up to exponential precision

Approximating γ using an encoded bitstream for x = 0

Recall that by Theorem 1.2.2, exponentially precise approximations $\tilde{\gamma}$ to the root γ in Theorem 1.2.4 are sufficient in order to reconstruct subsequent input $x_n = f(t_n) \in [-1, 1]$ whose bit streams are expansions in root γ with exponential precision. In this section, we present a method to approximate γ with such precision using the only information at our disposal at the decoding end of the quantization scheme: the encoded bitstreams of real numbers $x \in [-1, 1]$. More precisely, we will be able to recover the value $\gamma = \beta^{-1}$ using only a single bitstream corresponding to a beta-expansion of the number 0. It is easy to adapt this method to slow variations of γ in time, as one can repeat the following procedure at regular time intervals during quantization, and update the value of γ accordingly.

The analysis that follows will rely on the following theorem by Peres and Solomyak [61]:

Theorem 1.2.6 (Peres-Solomyak). *[δ-transversality] Consider the intervals $I_\rho = [0, \rho]$. If $\rho \leq .6491...,$ then for any g of the form*

$$g(x) = 1 + \sum_{j=1}^{\infty} b_j x^j, \ b_j \in \{-1, 0, 1\} \tag{1.17}$$

and any $x \in I_\rho$, there exists a $\delta_\rho > 0$ such that if $g(x) < \delta_\rho$ then $g'(x) < -\delta_\rho$.

Furthermore, as ρ increases to .6491..., δ_ρ decreases to 0.

Theorem 1.2.6 has the following straightforward corollary:

Corollary 1.2.7. *If $g(x)$ is a polynomial or power series belonging to the class B given by*

$$B = \{\pm 1 + \sum_{j=1}^{\infty} b_j x^j, \ : b_j \in \{-1, 0, 1\}\} \tag{1.18}$$

then g can have no more than one root on the interval $(0, .6491]$. Furthermore, if such a root exists, then this root must be simple.

In [61], Peres and Solomyak used Theorem 1.2.6 to show that the distribution ν_λ of the random series $\sum \pm \lambda^n$ is absolutely continuous for a.e. $\lambda \in (1/2, 1)$. The estimates in Theorem 1.2.6 are obtained by computing the smallest double zero of a larger class of power series $\bar{B} := \{1 + \sum_{n=1}^{\infty} b_n x^n : b_n \in [-1, 1]\}$. The specific upper bound $\rho = .6491...$ for which δ-transversality holds on the interval I_ρ is the tightest bound that can be reached using their method of proof, but the true upper bound cannot be much larger; in [68], a power series $g(t)$ belonging to the class B in (1.18) is constructed which has a double zero at $t_0 \approx .68$ (i.e., $g(t_0) = 0$ and $g'(t_0) = 0$), and having a double root obviously contradicts δ-transversality.

We now show how to use Theorem 1.2.6 to recover γ from a bitstream $(b_j)_{j=1}^{\infty}$ produced from the GRE scheme (1.16) corresponding to input $x = 0$. We assume that the parameters (α, ν) of the quantizer Q_α^ν used in this implementation are within the range provided by Theorem 1.2.5. Note that such a bitstream $(b_j)_{j=1}^{\infty}$ is not unique if $\nu > 0$, or if more than one pair (λ_1, λ_2) correspond to the same value of

36

γ. Nevertheless, Theorem 1.2.4 implies that $\sum_{j=1}^{\infty} b_j \gamma^j = 0$, so that γ is a root of the power series $F(t) = b_1 + \sum_{j=1}^{\infty} b_{j+1} t^j$. Suppose we know that $\beta \geq 1.54056$, or that $\gamma \leq .6491$. Since $F(t)$ belongs to the class B of Corollary (1.2.7), γ must necessarily be the *smallest* positive root of $F(t)$. In reality one does not have access to the entire bitsream $(b_j)_{j=1}^{\infty}$, but only a finite sequence $(b_j)_{j=1}^{N}$. It is natural to ask whether we can approximate the first root of $F(t)$ by the first root of the polynomials $P_n(t) = b_1 + \sum_{j=1}^{n} b_{j+1} t^j$. Since the P_n still belong to the class B of Corollary (1.2.7), $|P_n(t)|$ has *at most* one zero on the interval $[\phi^{-1}, .6491] \approx [.618, .6491]$. The following theorem shows that, if it is known a priori that $\gamma \leq .6491 - \epsilon$ for some $\epsilon > 0$, then for n sufficiently large, $|P_n(t)|$ is guaranteed to have a root γ_n in $[0, .6491]$, and $|\gamma - \gamma_n|$ decreases exponentially as n increases.

Theorem 1.2.8. *Suppose that for some $\epsilon > 0$, it is known that $\gamma \leq \gamma_{high} = .6491 - \epsilon$. Let $\delta > 0$ be such that δ-transversality holds on the interval $[0, \gamma_{high}]$. Let N be the smallest integer such that $\gamma^{N+1} \leq (1 - \gamma)\epsilon\delta$. Then for $n \geq N$,*

(a) P_n has a unique root γ_n in $[0, .6491]$

(b) $|\gamma - \gamma_n| \leq C_1 \gamma^n$, where $C_1 = \frac{\gamma}{\delta(1-\gamma)}$.

Proof. Without loss of generality, we assume $b_1 = 1$, and divide the proof into 2 cases: (1) $P_n(\gamma) \leq 0$, and (2) $P_n(\gamma) > 0$. The proof is the same for $b_1 = -1$, except that the cases are reversed.

Case (1): Suppose $P_n(\gamma) \leq 0$. In this case, many of the restrictions in the theorem are not necessary; in fact, the theorem holds here for all $n > 0$ and $\epsilon \geq 0$. $P_n(t)$ has opposite signs at $t = 0$ and $t = \gamma$, so P_n must have at least one root γ_n in between.

Moreover, this root is unique by Theorem 1.2.6. To prove part (b) of the theorem, observe that Theorem 1.2.6 implies that if $P_n(t_0) \leq \delta$ at some $t_0 \in (0, .6491)$, then $P_n'(t) \leq -\delta$ in the interval $[t_0, .6491)$. In particular, $P_n(t) \leq \delta$ and $P_n'(t) \leq -\delta$ for $t \in [\gamma_n, \gamma]$. By the Mean Value Theorem, $|P_n(\gamma)| = |P_n(\gamma) - P_n(\gamma_n)| = |P_n'(\xi)||\gamma - \gamma_n|$ for some $\xi \in [\gamma_n, \gamma]$, so that

$$
\begin{aligned}
|\gamma - \gamma_n| &\leq \frac{|P_n(\gamma)|}{\inf_{\xi \in [\gamma_n, \gamma]} |P_n'(\xi)|} \\
&\leq \frac{|P_n(\gamma)|}{\delta} \\
&\leq \frac{\gamma^{n+1}}{(1 - \gamma)\delta}.
\end{aligned}
$$

The inequality $|P_n(\gamma)| \leq \frac{\gamma^{n+1}}{1-\gamma}$ follows from Theorem 1.2.4.

Case (2): If $P_n(\gamma) > 0$, then Theorem 1.2.6 implies that the first positive root of P_n, if it exists, must be greater than γ. Whereas in Case (1), P_n was guaranteed to have a root $\gamma_n \leq \gamma$ for all $n \geq 0$, in this case P_n might not even have a root in $(0, 1)$; for example, the first positive root of the polynomial $P_4(t) = 1 - t - t^2 + t^3$ occurs at $t = 1$. However, if n is sufficiently large, P_n will have a root in $[\gamma, .6491]$. Precisely, let $n \geq N$, where N is the smallest integer such that $\gamma^{N+1} \leq (1 - \gamma)\epsilon\delta$. Then $P_n(\gamma) \leq \frac{\gamma^{n+1}}{1-\gamma} \leq \frac{\gamma^{N+1}}{1-\gamma} \leq \epsilon\delta \leq \delta$. Since $\gamma + \epsilon \leq .6491$, Theorem 1.2.6 gives that $P_n'(t) \leq -\delta$ for $t \in [\gamma, \gamma + \epsilon]$, and

$$
\begin{aligned}
P_n(\gamma + \epsilon) &\leq P_n(\gamma) + \epsilon \sup_{t \in [\gamma, \gamma + \epsilon]} P_n'(t) \\
&\leq \epsilon\delta - \epsilon\delta \\
&= 0.
\end{aligned}
$$

We have shown that $P_n(\gamma) > 0$ and $P_n(\gamma + \epsilon) \leq 0$, so P_n is guaranteed to have a root γ_n in the interval $[\gamma, \gamma + \epsilon]$, which is in fact the unique root of P_n in $[0, .6491]$. Furthermore, the discrepency between γ_n and γ becomes exponentially small as $n > N$ increases, as

$$|\gamma_n - \gamma| \leq \frac{|P_n(\gamma)|}{\inf_{\xi \in [\gamma, \gamma_n]} |P_n'(\xi)|} \leq \frac{\gamma^{n+1}}{(1-\gamma)\delta} .$$

$\square$

Remark 1.2.9. *In* $[61]$ *it is shown that δ-transversality holds on $[0, .63]$ with $\delta = \delta_{.63} = .07$. This interval corresponds to $\epsilon = .6491 - .63 = .0191$ in Theorem* 1.2.8. *If γ is known a priori to be less than .63, then $\frac{\gamma}{1-\gamma} \leq \frac{.63}{1-.63} \approx 1.7027$, and N of Theorem* (1.2.8) *satisfies*

$$N \leq \frac{\log \epsilon\delta - \log 1.7027}{\log .63} \leq 16.$$

Of course, even if we know $P_n(t)$, n will be too large to solve for γ_n analytically. However, if we can approximate γ_n by $\tilde{\gamma}_n$ with a precision of $O(\gamma^n)$, then $|\gamma - \tilde{\gamma}_n| \leq |\gamma - \gamma_n| + |\gamma_n - \tilde{\gamma}_n|$ will also be of order γ^n, so that the estimates $\tilde{\gamma}_n$ are still exponentially accurate approximations to the input signal.

Indeed, any $\tilde{\gamma}_n \in [\phi^{-1}, .6491 - \epsilon]$ satisfying $|P_n(\tilde{\gamma}_n)| \leq \phi^{-n}$ provides such an exponential approximation to γ_n. Since $n \geq N$, it follows that $\phi^{-n} \leq \gamma^n \leq \gamma^N \leq \frac{(1-\gamma)\epsilon\delta}{\gamma} \leq \delta$,

and so $P_n'(t) \leq -\delta$ on the interval between $\tilde{\gamma}_n$ and γ_n. Thus,

$$
\begin{aligned}
|\gamma - \tilde{\gamma}_n| \;&\leq\; |\gamma - \gamma_n| + |\gamma_n - \tilde{\gamma}_n| \\
&\leq\; C_1 \gamma^n + \frac{|P_n(\tilde{\gamma}_n)|}{\inf_{\xi \in [\tilde{\gamma}_n, \gamma_n] \, or \, \xi \in [\gamma_n, \tilde{\gamma}_n]} |P_n'(\xi)|} \\
&\leq\; C_1 \gamma^n + \frac{\phi^{-n}}{\delta} \\
&\leq\; C_1 \gamma^n + \frac{\gamma^n}{\delta} \\
&=\; C_2 \gamma^n,
\end{aligned}
\tag{1.19}
$$

where $C_2 = C_1 + \frac{1}{\delta} = \frac{1}{\delta(1-\gamma)}$.

Approximating γ with beta-expansions of $(-x, x)$

The method for approximating the value of γ in the previous section requires a bitstream b corresponding to running the GRE recursion (1.16) with specific input $x = 0$. This assumes that the reference value 0 can be measured with high precision, which is an impractical constraint. We can try to adapt the argument using bit-streams of an encoded pair $(x, -x)$ as follows. Let b and c be bitstreams correspond-ing to x and $-x$, respectively. Define $d_j = b_j + c_j$. Put $k = min\{j \in N | d_{j+1} \neq 0\}$, and consider the sequence $\bar{d} = (\bar{d}_j)_{j=1}^{\infty}$ defined by $\bar{d}_j = \frac{1}{2} d_{j+k}$. Since $d_j \in \{-2, 0, 2\}$,

it follows that $\bar{d}_j \in \{-1, 0, 1\}$, and by Theorem 1.2.4, we have that

$$
\begin{aligned}
\sum_{n=1}^{\infty} \bar{d}_n \gamma^n &= \frac{1}{2} \gamma^{-k} \sum_{n=1}^{\infty} d_n \gamma^n \\
&= \frac{1}{2} \gamma^{-k} \left[\sum_{n=1}^{\infty} b_n \gamma^n + \sum_{n=1}^{\infty} c_n \gamma^n \right] \\
&= 0 \tag{1.20}
\end{aligned}
$$

so that

$$
|\sum_{n=1}^{N} \bar{d}_n \gamma^n| \leq C_\gamma \gamma^N, \tag{1.21}
$$

where the constant $C_\gamma = \frac{\gamma}{1-\gamma}$.

Equation (1.21), along with the fact that the polynomials $\bar{P}_N(t) := \bar{d}_1 + \sum_{j=1}^{N} \bar{d}_{j+1} t^j$ are of the form (1.18), allows us to apply Theorem 1.2.8 to conclude that for N sufficiently large, the first positive root γ_N of $\bar{P}_N$ becomes exponentially close to γ. However, note that the encoding of $(x, -x)$ is *not* equivalent to the encoding of 0. The value of $k = min\{j \in N | d_{j+1} \neq 0\}$ used to define the sequence $\bar{d}$ can be *arbitrarily large*. In fact, if an ideal quantizer Q_α^0 is used, then the bitstreams b and c are uniquely defined by $b \equiv -c$, so that $d \equiv 0$. Thus, this method for recovering γ actually *requires* the use of a flaky quantizer Q_α^ν. To this end, one could intentially implement GRE with a quantizer which toggles close to, but not exactly at zero. One could alternatively send not only a single pair of bitstreams (b, c), but *multiple* pairs of bitstreams (b^l, c^l) corresponding to several pairs $(x_l, -x_l)$, to increase the chance of having a pair that has $b^{L^j} + c^{L^j} \neq 0$ for relatively small j.

Figure 1.2 plots several instances of $\bar{P}_8, \bar{P}_{16}$, and $\bar{P}_{32}$, corresponding to $\gamma = .64375$. The quantizer $Q_\alpha^\nu(u, v)$ is used, with $\alpha = 2$ and $\nu = .3$. These values of α and ν generate bounded sequences (u_n) for all $(\lambda_1, \lambda_2) \in [.9, 1]^2$ by Theorem 1.2.5. Figure 1.3 plots several instances of the same polynomials, but with root $\gamma = .75$.

As shown in Figure 1.1, numerical evidence suggests that 10 iterations of Newton's method starting from $x_0 = \phi^{-1} \approx .618$ will compute an approximation to γ_N, the first root of $\bar{P}_N$, with the desired exponential precision. The figure plots γ versus the error $|\gamma - \tilde{\gamma}_N|$, where $\tilde{\gamma}_N$ is the approximation to γ_N obtained via a 10- step Newton Method, starting from $x_0 = .618$. More precisely, for each N, we ran 100 different trials, with x and γ picked randomly from the intervals $[-1, 1]$ and $[.618, .7]$ respectively, and independently for each trial. The worst case approximation error of the 100 trials is plotted in each case. The quantizer used is $Q_\alpha^\nu(u, v)$ with $\nu = .3$, and α picked randomly in the interval $[1.7, 2]$, independently for each trial. Again, Theorem 1.2.5 shows that these values of α and ν generate bounded sequences (u_n) for all $(\lambda_1, \lambda_2) \in [.9, 1]^2$.

1.2.5 The beta-encoder revisited

Even though our analysis of the previous section was motivated by leaky GRE encoders, it can be applied to general beta-encoders to recover the value of β at any time during quantization. From the last section, we have:

Theorem 1.2.10. *Let* $F(t) = \sum_{j=0}^\infty b_j t^j$ *be a power series belonging to the class* $B = \{\pm 1 + \sum_{j=1}^\infty b_j x^j, \ b_j \in \{-1, 0, 1\}\}$. *Suppose that F has a root at $\gamma \in [\gamma_{low}, \gamma_{high}]$,*

42

where $\gamma_{high} = .6491 - \epsilon$ for some $\epsilon > 0$. Let $\delta > 0$ be such that δ-transversality holds on the interval $[0, \gamma_{high}]$. Let N be the smallest integer such that $\gamma^{N+1} \leq \epsilon\delta(1 - \gamma)$. Then for $n \geq N$,

(a) The polynomials $P_n(t) = \sum_{j=0}^{n} b_j t^j$ have a unique root γ_n in $[0, .6491]$

(b) Any $\tilde{\gamma} \in [\gamma_{low}, \gamma_{high}]$ which satisfies $|P_n(\tilde{\gamma})| \leq (\gamma_{low})^n$ also satisfies $|\tilde{\gamma} - \gamma| \leq \tilde{C}|\gamma|^n$, where $\tilde{C} = \frac{1}{\delta(1-\gamma)}$.

This theorem applies to beta encoders, corresponding to implementing the recursion (1.7) with flaky quantizer Q^ν defined by (1.9), and with $\gamma = \beta^{-1}$ known a priori to be contained in an interval $[\gamma_{low}, \gamma_{high}]$. If $\beta \geq 1.54059$, or $\gamma_{high} \leq .6491$, then we can recover $\gamma = \beta^{-1}$ from either a bitstream corresponding to 0, or a pair of bitstreams $(x, -x)$, using Theorem 1.2.10. Of course we should not consider only the scheme (1.7), but rather a revised scheme which accounts for integrator leak on the (single) integrator used in the beta-encoder implementation. The revised beta-encoding scheme, with slightly different initial conditions, becomes

$$
\begin{aligned}
u_1 &= x \\
b_1 &= Q(u_1) \\
\text{for } j \geq 1 : u_{j+1} &= \lambda\beta(u_j - b_j) \\
b_{j+1} &= Q(u_{j+1})
\end{aligned}
\tag{1.22}
$$

where λ is an unknown parameter in $[.9, 1]$. As long as $\tilde{\beta} = \lambda\beta > 1$, we still have that $|x - \sum_{i=0}^{N} b_{i+1}\gamma^i| \leq C_{\tilde{\beta}}\tilde{\beta}^{-N}$ where $C_{\tilde{\beta}} = \frac{1}{\tilde{\beta}-1}$; furthermore, we can use Theorem

43

1.2.10 to recover $\tilde{\beta} = \lambda\beta$ in (1.22), although the specific values of λ and β cannot be distinguished, just as the specific values of λ_1 and λ_2 in the expression for γ could not be distinguished in GRE.
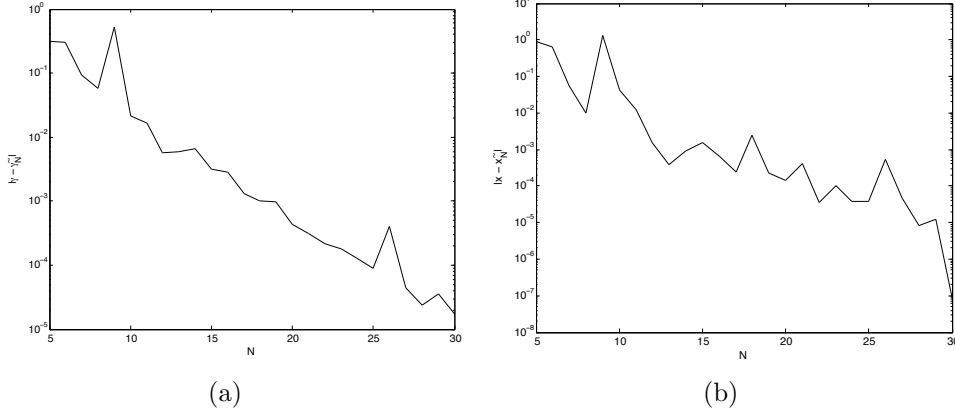

(a)   (b)

Figure 1.1: (a) N versus $|\gamma - \tilde{\gamma}_N|$, where $\tilde{\gamma}_N$ is obtained via a 10 step Newton Method approximation to the first root of the polynomial $\bar{P}_N$ starting from $x_0 = .618$. For each N, $|\gamma - \tilde{\gamma}_N|$ is the worst-case error among 100 experiments corresponding to different $(x_j, -x_j)$ pairs and different γ_j chosen randomly from $[.618, .7]$. (b) N versus $|x - \tilde{x}_N|$, where $\tilde{x}_N$ is reconstructed from $\tilde{\gamma}_N$ using the first N bits. The quantizer used in this experiment is Q_α^ν with $\nu = .3$ and α chosen randomly in the interval $[1.7, 2]$.

Remarks

Figure 1.1 suggests that the first positive roots of the P_n serve as exponentially precise approximations to γ for values of γ greater than .6491; Figure 1.3 suggests that the first positive root of P_n will approximate values of γ up to $\gamma = .75$. Furthermore, these figures suggest that the constants C_1 and C_2 of (1.19) in the exponential convergence of these roots to γ can be made much sharper, even for larger values of γ. This should not be surprising, considering that nowhere in the analysis of the

44

previous section did we exploit the specific structure of beta-expansions obtained via the particular recursions (1.22) and (1.12), such as the fact that such sequences (b_n) cannot contain infinite strings of successive 1's or -1's. It is precisely power series with such infinite strings that are the 'extremal cases' which force the bound of .6491 in Theorem 1.2.6. It is difficult to provide more refined estimates for the constants C_1 and C_2 of (1.19) in general, but in the idealized setting where beta-expansions of 0 are available via the ideal GRE scheme (1.12) without leakage, or via the beta-encoding (1.12) with $\beta = \phi$, the beta-expansions of 0 have a very special structure:

Proposition 1.2.11. *Consider the ideal GRE recursion (1.12) with input $u_0 = u_1 = 0$ and $Q = Q_\alpha^\nu$, or the beta-encoder recursion (1.7) with $\beta = \phi$, $u_0 = 0$, and $Q = Q^\nu$. As long as $\alpha > \nu$ in (1.12), or $\nu \leq 1$ in (1.7), then for each $j \in \{0, 1, ...\}$, b_{3j} is equal to -1 or $+1$, and $b_{3j+1} = b_{3j+2} = -b_{3j}$.*

The proof is straightforward, and we omit the details.

Proposition 1.2.11 can be used to prove directly that γ must be the first positive root of the polynomials $P_N(t) = b_1 + \sum_{j=1}^{3N} b_{j+1} t^j$, when $\gamma = \phi^{-1}$ in either (1.12) or (1.7). Indeed, P_N can be factored as follows:

$$
\begin{aligned}
P_N(t) &= \sum_{j=0}^{N} (b_{3j+1} - b_{3j+1} t - b_{3j+1} t^2) t^{3j} \\
&= (1 - t - t^2) \sum_{j=1}^{N} b_{3j+1} t^{3j} \\
&= (1 - t - t^2) R_N(t) \qquad\qquad (1.23)
\end{aligned}
$$

45

where R_N is a polynomial with random coefficients of the form $\sum_{j=0}^{N} \pm t^{3j}$.

$P_N(t) = (1 - t - t^2)R_N(t)$ clearly has a root at $t = \phi^{-1}$, and this root must be the only root of P_N on $[0, \phi^{-1})$, since on this interval R_N is bounded away from 0 by $|R_N(t)| \geq 1 - \frac{t^3}{1-t^3} \geq 1 - \frac{\phi^{-3}}{1-\phi^{-3}} \approx .691$. We can also obtain a lower bound on $|P_N'(\phi^{-1})| = -(1 + 2\phi^{-1})R_N(\phi^{-1})$ by $|P_N'(\phi^{-1})| \geq |1 + 2\phi^{-1}||1 - \frac{\phi^{-3}}{1-\phi^{-3}}| \geq 1.545$. Note that this bound holds also for the infinite sum $F(t) = b_1 + \sum_{j=1}^{\infty} b_{j+1}t^j$, and that the bound of $|F'(\gamma)| \geq 1.545$ is much sharper than the bound on $|F'(\gamma)|$ given by Theorem 1.2.6; e.g., $\delta_{.63} = .07$, and $\delta_{.6491} = .00008$ (see [61]). Similar bounds on the derivatives $|P_N'(\gamma)|$ and $|F_N'(\gamma)|$ corresponding to beta-expansions of 0 in a base γ close to ϕ^{-1} should hold, leading to sharper estimates on the constants C_1 and C_2 of (1.19) in the case where beta-expansions of 0 are available.

1.2.6 Closing remarks

We have shown that golden ratio encoders are robust with respect to leakage in the delay elements of their circuit implementation. Although such leakage may change the base γ in the reconstruction formula $|y - \sum_{j=1}^{N} b_j \gamma^j| \sim O(\gamma^N)$, we have shown that exponentially precise approximations $\tilde{\gamma}$ to γ can be obtained from the bitstreams of a pair $(x, -x)$, and such approximations $\tilde{\gamma}$ are sufficient to reconstruct subsequent input y by $|y - \sum_{j=1}^{N} b_j \tilde{\gamma}^j| \sim O(\gamma^N)$.

Our method can be extended to recover the base β in general beta-encoders, as long as β is known a priori to be sufficiently large; e.g. $\beta \geq 1.54$. This method

is similar to the method proposed in [33] for recovering β in beta-encoders when $\{0, 1\}$-quantizers are used, except that our method does not require a fixed reference level, which is difficult to measure with high precision in practice.
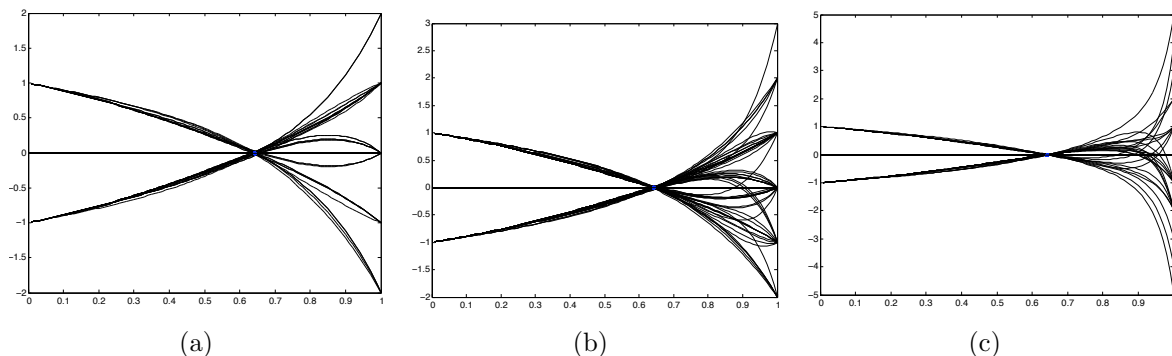

Figure 1.2: The graphs of (a) $\bar{P}_8$, (b) $\bar{P}_{16}$, and (c) $\bar{P}_{32}$ for several pairs $(x_j, -x_j)$ corresponding to $\gamma = .64575$. These polynomials can have several roots on the unit interval, but the first of these roots must become exponentially close to γ as N increases. The quantizer used is Q_α^ν with parameter values $\nu = .3$, and $\alpha = 2$.

1.2.7   Appendix: proof of Theorem 1.2.5

In this section we prove Theorem 1.2.5, which provides a range within which the "flakiness" parameter ν and the "amplifier" parameter α in the quantizer $Q_\alpha^\nu(u, v)$ can vary from iteration to iteration, without changing the fact that the sequences (u_n) produced by the scheme (1.16) will be bounded. The derived range for ν and α is independent of the specific values of the parameters $(\lambda_1, \lambda_2) \in [.9, 1]^2$ in (1.16).

The techniques of this section are borrowed in large part from those used in [35] to prove a similar result for the ideal GRE scheme (1.12); i.e., taking $\lambda_1 = \lambda_2 = 1$
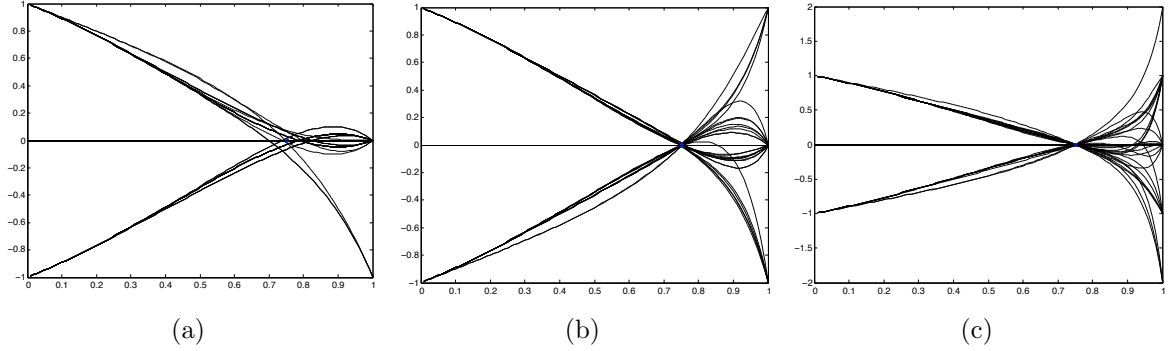
Figure 1.3: The graphs of (a) $\bar{P}_8$, (b) $\bar{P}_{16}$, and (c) $\bar{P}_{32}$ for several pairs $(x_j, -x_j)$ corresponding to $\gamma = .75$. These graphs suggest that the first positive root of these polynomials becomes exponentially close to γ as N increases, although we do not have a proof of this result for values of γ greater than approximately .65. The quantizer used is Q_α^ν with parameter values $\nu = .3$, and $\alpha = 2$.
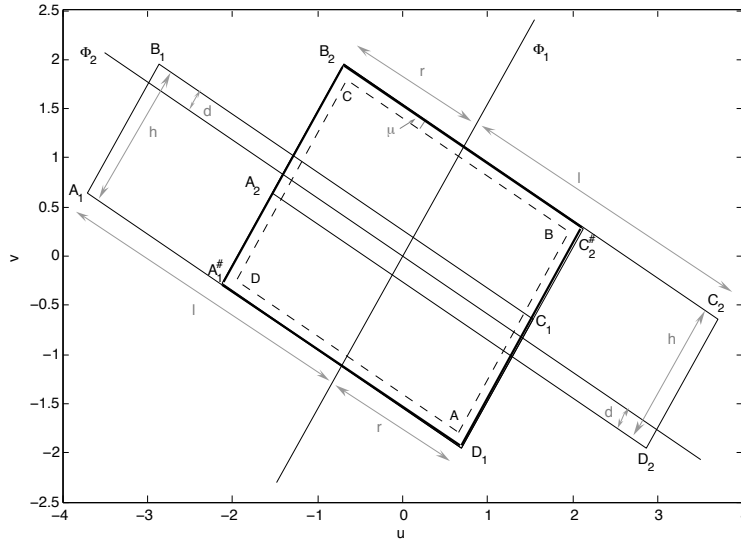
Figure 1.4: The rectangle $R = A_1^{\#} B_2 C_2^{\#} D_1$ (bold outline) is a positively invariant set for the map T_α^ν. In fact, $T_\alpha^\nu(R)$ is contained in the smaller rectangle $ABCD$ (dashed outline), illustrating that the revised GRE scheme (1.16) is robust with respect to small additive errors. Here, Φ_1 and Φ_2 are the eigenvectors of the matrix in A in (1.28) below. In this figure, $\lambda_2 = .6$ and μ is taken to be .0625. The length parameters l, r, h, and d will be discussed later.

48

in (1.25). As is done in [35], we first observe that the recursion formula (1.16) corresponding to the leaky GRE is equivalent to the following piecewise affine discrete dynamical system on $\mathbb{R}^2$:

$$\begin{bmatrix} u_{n+1} \\ u_{n+2} \end{bmatrix} = T_\alpha^\nu \begin{bmatrix} u_n \\ u_{n+1} \end{bmatrix} \tag{1.24}$$

where

$$T_\alpha^\nu : \begin{bmatrix} u \\ v \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 1 \\ \lambda_1\lambda_2 & \lambda_1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} - Q_{\tilde{\alpha}}^{\tilde{\nu}}(u,v) \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{1.25}$$

where $\tilde{\alpha} = \frac{\alpha}{\lambda_2}$ and $\tilde{\nu} = \frac{\nu}{\lambda_1\lambda_2}$. We will construct a class of subsets $R = R(\mu) = R(\lambda_1, \lambda_2, \mu)$ of $\mathbb{R}^2$ for which $T_\alpha^\nu(R) + B_\mu(0) \subset R$, where $B_\mu(0)$ is the disk of radius μ centered at the origin; i.e. $B_\mu(0) = \{(u,v) : u^2 + v^2 \leq \mu^2\}$. Note that these sets are not only *positively invariant sets* of the map T_α^ν, but have the additional property that if $(u_0, v_0) \in R(\mu)$, the image $(u_{n+1}, v_{n+1}) = T_\alpha^\nu(u_n, v_n)$ may be perturbed at any time within a radius of μ (for example, by additive noise), and the the resulting sequence $(u_n)_{n=0}$ will still remain bounded within $R(\mu)$ for all time n.

We refer the reader to Figure 1.4 as we detail the construction of the sets $R(\mu)$. Rectangles $A_1B_1C_1D_1$ and $A_2B_2C_2D_2$ in Figure 1.4 are designed so that their respective images under the affine maps T_1 and T_2 (defined below) are both equal to

49

the dashed rectangle $ABCD$.

$$T_1 : \begin{bmatrix} u \\ v \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 1 \\ \lambda_1 \lambda_2 & \lambda_1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$T_2 : \begin{bmatrix} u \\ v \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 1 \\ \lambda_1 \lambda_2 & \lambda_1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{1.26}$$

That is, rectangles $A_1 B_1 C_1 D_1$ and $A_2 B_2 C_2 D_2$ will satsify $T_1(A_1 B_1 C_1 D_1) = ABCD = T_2(A_2 B_2 C_2 D_2)$. More specifically, $T_1(A_1) = T_2(A_2) = A$, $T_1(B_1) = T_2(B_2) = B$, and so on. Since the rectangle $R = A_1^{\#} B_2 C_2^{\#} D_1$ is contained within the union of $A_1 B_1 C_1 D_1$ and $A_2 B_2 C_2 D_2$, R is a positively invariant set for any map $T(u, v)$ satisfying

$$T(u, v) = \begin{cases} T_1(u, v), & (u, v) \in R \setminus A_2 B_2 C_2 D_2 \\ T_2(u, v), & (u, v) \in R \setminus A_1 B_1 C_1 D_1 \\ T_1(u, v) \; or \; T_2(u, v), & (u, v) \in H \end{cases} \tag{1.27}$$

where $H = A_1 B_1 C_1 D_1 \cap A_2 B_2 C_2 D_2$. In particular, if the parameters α and ν in the map T_α^ν are chosen such that the intersection of R and the strip $F = \{(u, v) : -\tilde{\nu} < u + \tilde{\alpha} v < \tilde{\nu}\}$ is a subset of H (recall $\tilde{\alpha} = \alpha/\lambda_2$ and $\tilde{\nu} = \nu/(\lambda_1 \lambda_2)$), then T_α^ν is of the form (1.27). Indeed, F is the region of the plane in which the quantizer $Q_{\tilde{\alpha}}^{\tilde{\nu}}(u, v)$ operates in flaky mode. This geometric setup is clarified with a figure, provided in Figure 1.2.7.

It remains to verify the existence of at least one solution to the setup in Figure 1.4 for each $(\lambda_1, \lambda_2) \in [.9, 1]^2$. This can be done, following the lead of [35], in terms of

50
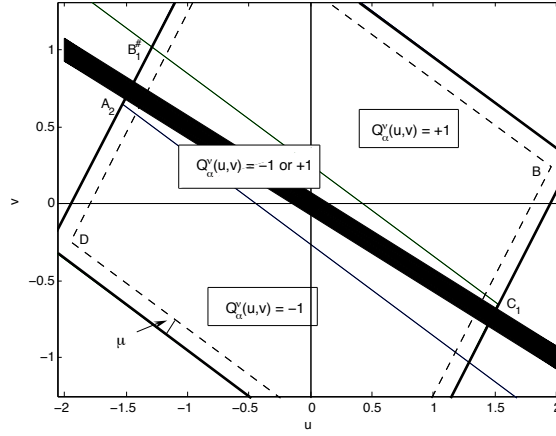
Figure 1.5: *The flaky quantizer $Q_{\tilde{\alpha}}^{\tilde{\nu}}(u, v)$ outputs either -1 or $+1$ when the input (u, v) belongs to the strip $F = \{(u, v) : \tilde{\nu} < u + \tilde{\alpha}v < \tilde{\nu}\}$ (shaded above). Here $\mu = .0625$, $\tilde{\alpha} = 2$, and $\tilde{\nu} = .15$. For these parameter values, the intersection of F and R is a subset of $A_1 B_1 C_1 D_1 \cap A_2 B_2 C_2 D_2$, meaning that the scheme (1.16) produces bounded sequences (u_n), when implemented with this quantizer, and with input $(u_0, u_1) \in R$.*

the parameters defined in Figure 1.4.

Note that the matrix

$$A = \begin{bmatrix} 0 & 1 \\ \lambda_1 \lambda_2 & \lambda_1 \end{bmatrix} \tag{1.28}$$

in (1.26) has as eigenvalues $\epsilon_1 = \frac{\lambda_1 + \sqrt{\lambda_1^2 + 4\lambda_1\lambda_2}}{2}$ and $-\epsilon_2 = -\left(\frac{-\lambda_1 + \sqrt{\lambda_1^2 - 4\lambda_1\lambda_2}}{2}\right)$. In particular, when $\lambda_1 = \lambda_2 = 1$, we have that $\epsilon_1 = \phi \approx 1.618$ and $\epsilon_2 = \phi^{-1} \approx .618$. The eigenvalues ϵ_1 and $-\epsilon_2$ have respective normalized eigenvectors

$$\Phi_1 = s_1^{-1} \begin{bmatrix} 1 \\ \epsilon_1 \end{bmatrix} \text{ and } \Phi_2 = -s_2^{-1} \begin{bmatrix} 1 \\ \epsilon_2 \end{bmatrix},$$

where $s_1 = \sqrt{1 + \epsilon_1^2}$ and $s_2 = \sqrt{1 + \epsilon_2^2}$. It follows that the affine map T_1 acts as an expansion by a factor of ϵ_1 along Φ_1 and a reflection followed by contraction by a factor of ϵ_2 along Φ_2, followed by a vertical translation of $+1$. T_2 is the same as T_1 except with a vertical translation of -1 instead of +1. After some straightforward algebraic calculations, the mapping relations described above imply that the parameters in Figure 1.4 are given by

$$\begin{aligned} h &= \frac{2\mu}{1 - \epsilon_1} + \frac{2s_1}{\epsilon_1(\epsilon_1 - 1)(\epsilon_1 + \epsilon_2)} \\ d &= \frac{\mu}{1 - \epsilon_1} + \frac{s_1(2 - \epsilon_1)}{\epsilon_1(\epsilon_1 - 1)(\epsilon_1 + \epsilon_2)} \\ l &= \frac{\mu}{1 - \epsilon_2} + \frac{s_2(2 - \epsilon_2)}{\epsilon_2(1 - \epsilon_2)(\epsilon_1 + \epsilon_2)} \\ r &= \frac{\mu}{1 - \epsilon_2} + \frac{s_2}{(1 - \epsilon_2)(\epsilon_1 + \epsilon_2)} \end{aligned} \tag{1.29}$$

The existence of the overlapping region $H = A_1 B_1 C_1 D_1 \cap A_2 B_2 C_2 D_2$ is equivalent to the condition $d > 0$ which in turn is equivalent to the condition $\mu < \frac{s_1(2-\epsilon_1)}{\epsilon_1(\epsilon_1+\epsilon_2)}$. This expression is minimized over the range $(\lambda_1, \lambda_2) \in V = [.9, 1]^2$ when $\lambda_1 = \lambda_2 = 1$, in which case this constraint simplifies to $\mu < \frac{2-\phi}{\sqrt{\phi^2+1}} \approx .2008$.

We are interested primarily in quantizing numbers in $[-1, 1]$, so that this is the range of interest for the input $u_1 = x$; for u_0 we simply take 0. The set $\{0\} \times [-1, 1] \subset \bigcap_{(\lambda_1, \lambda_2) \in V} R(\lambda_1, \lambda_2, \mu)$ is equivalent to the condition that the line $v = mu + b$ which passes through the points B and C in Figure 1.4 have y-intercept $b \geq 1$ for all $(\lambda_1, \lambda_2) \in V$. This line has slope $m = -\epsilon_2$, and passes through the point C, so that its y-intercept is given by $b = s_1^{-1}(\epsilon_1 + \epsilon_2)(h - d)$. Rearranging terms, this implies that $b \geq 1$ if and only if

$$\mu \leq \frac{(\sqrt{1 + \epsilon_1^2})(2 - \epsilon_1)}{\epsilon_1 + \epsilon_2}. \tag{1.30}$$

The right hand side of the above inequality is bounded below over the range $(\lambda_1, \lambda_2) \in V$ by $\frac{\sqrt{1+(.9\phi)^2}(2-\phi)}{\sqrt{5}} \approx .301$, so that indeed $\{0\} \times [-1, 1] \subset R(\lambda_1, \lambda_2, \mu)$ is satisfied independent of $(\lambda_1, \lambda_2) \in V$, for all admissable $\mu \in [0, .2008]$.

In practice, the "flaky" parameter ν of the quantizer Q_α^ν is not known exactly; we will however be given a tolerance δ for which it is known that $|\nu| \leq \delta$. It follows that for each δ, we would like a range $(\alpha_{min}, \alpha_{max})$ for the "amplifier" α such that the GRE scheme (1.16), implemented with quantizer Q_α^ν, produces bounded sequences (u_n) for all values of $(\lambda_1, \lambda_2) \in V$. Now, it will be easier to derive all of the following results for $\tilde{\alpha}$ and $\tilde{\nu}$, and return to the original parameters (α, ν) afterwards. For a

particular choice of (λ_1, λ_2), it is not hard to derive an admissable range for $\tilde{\alpha}$ in terms of the eigenvalues ϵ_1 and ϵ_2 for fixed δ. We will take $\mu = 0$ in the following analysis for the sake of simplicity. First of all, the tolerance δ must be admissable, i.e. the coordinate $(\delta, 0)$ should lie within the region H. Indeed, for $\tilde{\nu} \in [0, \delta]$, $\tilde{\alpha}$ can vary within a small neighborhood of $\frac{1}{\epsilon_2}$, as long as the line with slope $-\frac{1}{\tilde{\alpha}}$ which passes through $(0, \delta)$ remains bounded above in the region H by the line passing through B_1 and C_1. In other words, the admissable range for $\tilde{\alpha}$ is obtained from the constraint $m_1 \le -\frac{1}{\tilde{\alpha}} \le m_2$, where m_1 is the slope of the line through the points $\{(\delta, 0), C_1\}$ and m_2 is the slope of the line through the points $\{(\delta, 0), B_1^{\#}\}$ in Figure 1.2.7. Rewriting C_1 and $B_1^{\#}$ in terms of the eigenvalues ϵ_1 and ϵ_2 via the relations (1.29), we have that

$$L(\epsilon_1, \epsilon_2) \le \tilde{\alpha} \le U(\epsilon_1, \epsilon_1 + \epsilon_2) \tag{1.31}$$

where

$$
\begin{aligned}
L(x, y) &= N(x, y)/D(x, y) \\
&= \frac{x(x-1) - (2-x)(1-y) + \delta x(x-1)(x+y)(1-y)}{x((2-x)(1-y) + y(x-1))}
\end{aligned}
$$

and

$$U(x, y) = \frac{2 + xy - 2y - \delta xy(x-1)(1-y+x)}{x(y-2)}. \tag{1.32}$$

It is not hard to show that for any fixed $y \in [.9\phi + .9\phi^{-1}, \phi + \phi^{-1}] \approx [2.0124, 2.236]$, the function $U(x, y)$ increases as a function of x over the range $x \in [.9\phi, \phi] \approx [1.456, 1.618]$, as long as $\tilde{\nu} \le .75$. It follows that the minimum of $U(x, y)$ over

the rectangle $S = [1.456, 1.618] \times [2.0124, 2.236]$ occurs along the edge corresponding to $x = 1.456$. But $U(1.456, y)$ decreases as a function of y over the interval $[2.0124, 2.236]$, so that the minimum of $U(x, y)$ over the entire rectangle S occurs at $(x, y) = (1.456, 2.236)$, giving the following uniform lower bound on $\tilde{\alpha}_{max}$ for $(\lambda_1, \lambda_2) \in V$:

$$
\begin{aligned}
\tilde{\alpha}_{max} &\geq U(1.456, 2.236) \\
&\approx 2.281 - .952\delta.
\end{aligned}
\tag{1.33}
$$

We proceed in the same fashion to derive a uniform upper bound on $\tilde{\alpha}_{min}$ over $(\lambda_1, \lambda_2) \in V$, except that we analyze the numerator and the denominator in the expression for $L(x, y)$ separately. The numerator $N(x, y)$ increases as a function of x over the range $x \in [1.456, 1.618]$ for fixed $y \in [.9\phi^{-1}, \phi^{-1}] \approx [.5562, .618]$, so that $N(x, y)$ achieves its maximum over the rectangle $T = [1.456, 1.618] \times [.5562, .618]$ along the edge corresponding to $x = 1.618$. $N(1.618, y)$ increases as a function of y over the range $y \in [.5562, .618]$, so that $N(x, y)$ achieves its maximum over T at $(x, y) = (1.618, .618)$. A similar analysis shows that the denominator $D(x, y)$ is minimized over T at $(x, y) = (1.456, .618)$. It follows that for $(\lambda_1, \lambda_2) \in [.9, 1]^2$, $\tilde{\alpha}_{min}$ is bounded above by

$$
\begin{aligned}
\tilde{\alpha}_{min} &\leq N(1.618, .618)/D(1.456, .618) \\
&\approx 1.198(1 + \delta)
\end{aligned}
\tag{1.34}
$$

Finally, we note that δ is admissable for all (λ_1, λ_2) if and only if $L(\epsilon_1, \epsilon_2) \leq$

$U(\epsilon_1, \epsilon_1 + \epsilon_2)$ for this value of δ and for all coresponding ϵ_1 and ϵ_2. Thus an admissable range for δ is obtained by equating the lower bound of $2.281 - .952\delta$ in (1.33) and the upper bound of $1.198(1 + \delta)$ in (1.34); namely, $\delta \in [0, .5037]$.

In terms of the original parameters $\alpha = \lambda_2 \tilde{\alpha}$ and $\nu = \lambda_1 \lambda_2 \tilde{\nu}$),

$$
\begin{aligned}
1.198(1 + \delta) & \leq & \alpha \leq 2.053 - .8568\delta, \text{ and so} \\
0 & \leq & \delta \leq .4161, \\
\nu & = & \lambda_1 \lambda_2 \tilde{\nu} \leq \lambda_1 \lambda_2 \delta \leq .337.
\end{aligned}
\tag{1.35}
$$

represents a uniformly admissible range for (α, ν) over the interval $(\lambda_1, \lambda_2) \in [.9, 1]^2$.

Chapter 2

Quiet quantization in analog to digital conversion

In the last chapter, we studied robustness properties of the β-encoder, a quantization method in analog to digital conversion that is both robust with respect to quantizer imperfections like $\Sigma\Delta$ schemes, and also achieves optimal reconstruction guarantees like pulse code modulation (PCM). While we were able to show that β-encoders are also robust with respect to other component imperfections, the β-encoder is *not* robust with respect to additive noise; that is, if the discrete input x_n to the β-encoder (1.7) represent noisy signal samples $x_n = f_n + \epsilon_n$, then unavoidably

$$\sum_{j=0}^{N-1} b_j \beta^{-j} = f_n + \epsilon_n + O(\beta^{-N}), \tag{2.1}$$

and the situation is similar for the golden ratio encoder. That is, no matter how many bits are spent per Nyquist interval, the reconstruction error of the beta-encoder and golden ratio encoder is limited by the level of noise on the samples f_n.

On the other hand, $\Sigma\Delta$ schemes *are* robust with respect to additive noise. That is, the $O(\lambda^{-K})$ reconstruction accuracy of the Kth order $\Sigma\Delta$ scheme is preserved under additive noise, as facilitated by the incorporation of all previous samples f_k^λ, $k \leq n$, in determining the quantization output b_n. However, $\Sigma\Delta$ schemes suffer from an entirely different kind of instability in the context of audio signal processing, corresponding to a mechanical rather than mathematical reconstruction inaccuracy: at the onset of periodicities in the bit output b_n, the filters used in the reconstruction of the analog signal by $\Sigma\Delta$ quantizers are such that periodic oscillatory patterns in the quantization output b_n generate low-level *idle tones* that are not present in the signal. These tones are particularly intolerable to the listener when no signal is present, such as between successive audio tracks. As idle tone components are most prominent in low-order, 1-bit $\Sigma\Delta$ schemes [5], it is desirable to adapt these schemes in such a way as to eliminate the possibility of periodic behavior of the output b_n for vanishing input x_n; we shall refer to such as *quiet $\Sigma\Delta$ quantizers*, in line with the following definition:

Definition 2.0.12 (**Quiet quantization**). *A quantization scheme is said to be quiet if the onset of identically zero input $x_n \equiv 0$ after some finite time k induces constant bit output $b_n \equiv b_m$ after some later time $m \geq k$.*

In the first order $\Sigma\Delta$ model (2.2), which, starting from initial state u_0, follows the recursion

$$
\begin{aligned}
b_n &= Q(u_{n-1} + f_n^\lambda) \\
u_n &= u_{n-1} + f_n^\lambda - b_n,
\end{aligned}
\tag{2.2}
$$

one can eliminate periodicities in the output b_n at instantiation of small input f_n by simply replacing the standard 2-level quantizer $Q(u) = \text{sgn}(u)$ by a *tri-level* quantizer Q_{tri}^τ:

$$
Q_{tri}^\tau(u) = \begin{cases}
1 & u > \tau, \\
0 & -\tau \leq u \leq \tau, \\
-1 & u < -\tau.
\end{cases}
\tag{2.3}
$$

It is straightforward that the first-order $\Sigma\Delta$ scheme retains its first-order approximation accuracy when implemented with the tri-level quantizer $Q_{tri}^{.5}$, and is also quiet, in the sense of Definition 2.0.12; however, first order schemes are rarely used in practice, and we would thus like a similar result for higher order schemes. Unfortunately,

the second-order $\Sigma\Delta$ scheme,

$$
\begin{aligned}
b_n &= Q(F(u_{n-1}, v_{n-1})) \\
u_n &= u_{n-1} + f_n^\lambda - b_n \\
v_n &= v_{n-1} + u_n,
\end{aligned}
\tag{2.4}
$$

is not quiet, as shown in [79], even when implemented with tri-level quantizer Q_{tri}^τ and with linear rule $F_\gamma(u, v) = u + \gamma v$. In fact, for most initial states (u_0, v_0) at the onset of zero input $f_1^\lambda = 0$, the sequence b_n will become oscillatory. An exception to this rule occurs for the initial condition $(u_0, v_0) = (0, 0)$, in which case $(u_n, v_n) = (0, 0)$ and $b_n = 0$ persist as long as $f_n^\lambda = 0$ remains zero. That is, the origin is a fixed point of the *zero-input* discrete map $M : (u_n, v_n) \to (u_{n+1}, v_{n+1})$ produced by (2.4) with tri-level quantizer and zero input $f_n^\lambda = 0$, but is not an *attractive* fixed point. Recall that an attractive fixed point for a discrete map M is a fixed point $x_0 = M(x_0)$ having the property that for any value of x in the domain that is close enough to x_0, the *orbit* of x,

$$
(x, M(x), M(M(x)) = M^2 x, ...)
$$

converges to x_0. Such a fixed point is said to be a *global* attracting fixed point if $M^n x \to x_0$ for any value of x in the domain. In order to modify the second-order scheme (2.4) so that the origin is an attractive fixed point for the corresponding zero-input map, the state sequence (u_n, v_n) must be made to *decay*. One way to induce decay is by applying a contraction $(u, v) \to (\rho u, \rho v)$ before each iteration of

60

(2.4), resulting in the modification

$$
\begin{aligned}
b_n &= Q(F(\rho u_{n-1}, \rho v_{n-1})) \\
u_n &= \rho u_{n-1} + x_n - b_n, \\
v_n &= \rho v_{n-1} + u_n.
\end{aligned}
\tag{2.5}
$$

This so-called *finite-memory* scheme was first studied in [79]. As the accuracy of $\Sigma\Delta$ methods rely on 'keeping track' of the previous input $f_k^\lambda, k \leq n$ through the sequence (u_n, v_n), it is not clear that the finite-memory scheme, which loses a fraction of its 'memory' at each step, should still be second-order. However, as long as $1 - \rho$ is sufficiently small, second-order accuracy *is* maintained, as proven in [79]. For completeness, we include this result below, starting with the analogous result for the first-order finite-memory scheme,

$$
\begin{aligned}
b_n &= Q(\rho u_{n-1} + f_n^\lambda), \\
u_n &= \rho u_{n-1} + f_n^\lambda - b_n.
\end{aligned}
\tag{2.6}
$$

We note that the following result does not depend on the choice of quantizer, and requires only that the sequence (u_n) be uniformly bounded.

Lemma 2.0.13 (Yilmaz). *Suppose $\lambda > 0$, let $f \in L^2(\mathbb{R})$ be bandlimited with supp $f \subset [-\pi, \pi]$ and $\|f\|_\infty \leq 1$, and let g be a function satisfying $\hat{g} = 1$ for $|\xi| \leq \pi$, $\hat{g} = 0$ for $|\xi| \geq \lambda\pi$, and $g \in C^\infty$. Suppose that the leakage factor $\rho \geq \rho^\lambda = 1 - \frac{1}{\lambda}$, and assume that the sequence (u_n) generated by the finite-memory scheme (2.6) is*

61

bounded. If (b_n^λ) is the output of the first-order finite $\Sigma\Delta$-quantizer (2.6), then

$$\|f(t) - \tilde{f}(t)\|_\infty \le \frac{\|v\|_{\ell_\infty}}{\lambda}(\|g'\|_{L^1} + C_g),\tag{2.7}$$

where $C_g \le \|g\|_{L^1} + \frac{1}{\lambda}\|g'\|_{L^1}$, and $\tilde{f}(t) = \frac{1}{\lambda}\sum b_n^\lambda g(t - \frac{n}{\lambda})$.

Proof. We have $u_n - \rho u_{n-1} = f_n^\lambda - b_n^\lambda$. Therefore,

$$
\begin{aligned}
f(t) - \tilde{f}(t) &= \frac{1}{\lambda}\left(\sum(u_n - u_{n-1})g(t - \frac{n}{\lambda}) + (1 - \rho)\sum u_{n-1}g(t - \frac{n}{\lambda})\right) \\
&= \frac{1}{\lambda}\left(\sum u_n\big(g(t - \frac{n}{\lambda}) - g(t - \frac{n+1}{\lambda})\big) + (1 - \rho)\sum u_n g(t - \frac{n}{\lambda})\right).
\end{aligned}
$$

Using the bound $1 - \rho \le \frac{1}{\lambda}$,

$$
\begin{aligned}
|f(t) - \tilde{f}(t)| &\le \frac{\|u\|_{\ell_\infty}}{\lambda}\left(\sum|g(t - \frac{n}{\lambda}) - g(t - \frac{n+1}{\lambda})| + \frac{1}{\lambda}\sum|g(t - \frac{n}{\lambda})|\right) \\
&\le \frac{\|u\|_{\ell_\infty}}{\lambda}(\|g'\|_{L^1} + C_g).
\end{aligned}
$$

$\square$

The analogous result for the second-order scheme follows the same argument, and we state the result without proof.

Lemma 2.0.14 (Yilmaz). *Let f, g, and ρ be as in Lemma 2.0.13. Assume that the sequence (u_n, v_n) generated by (2.5) is bounded. If (b_n^λ) is the output of the second-order leaky $\Sigma\Delta$-quantizer given in (2.5), then*

$$|f(t) - \tilde{f}(t)| \le \frac{\|v\|_{\ell_\infty}}{\lambda^2}(\|g''\|_{L^1} + 2\|g'\|_{L^1} + 2C_g),\tag{2.8}$$

where C_g and $\tilde{f}$ are as before.

It is important to note that Lemma 2.0.14 is useful only in the finite regime, even though it is asymptotic statement. Indeed, the oversampling ratio λ is not taken to be arbitrarily large in practice as limited by the resolution of analog-to-digital technologies; in most A/D converters, $\lambda \leq 64$ is set. For this range of λ, the assumption in Lemma 2.0.14 that the damping factor ρ^λ can be prescribed up to resolution $\rho \in [.97, 1]$ is certainly reasonable.

Lemma 2.0.14 rests on the assumption that the initial input (u_0, v_0) can be prescribed so that the sequence (u_n, v_n) remains uniformly bounded. We can ensure such stability by constructing positively invariant sets for the discrete map $(u_n, v_n) \rightarrow (u_{n+1}, v_{n+1})$ corresponding to the finite-memory scheme (2.5). Recall that a set S is positively invariant for the discrete map M if $x \in S$ implies that $M(x) \in S$. The following theorem, borrowed again from [79], gives explicit constructions of such positively invariant sets for the finite-memory map (2.5) with linear rule $F_\gamma(u, v) = u + \gamma v$:

Theorem 2.0.15 (Yilmaz). *Fix $\alpha < 1$, and constant C satisfying*

$$C \geq \frac{1}{2 - 2\alpha^2} + \frac{12 + 9(1 + \alpha)}{8(1 - \alpha)}.$$

Consider the following functions,

$$B_1(u) = \begin{cases} -\frac{u^2}{2(1-\alpha)} + \frac{u}{2} + C, & u \geq 0 \\ -\frac{u^2}{2(1+\alpha)} + \frac{u}{2} + C, & u < 0 \end{cases} \tag{2.9}$$

$$B_2(u) = \begin{cases} \frac{u^2}{2(1+\alpha)} + \frac{u}{2} - C, & u \geq 0 \\ \frac{u^2}{2(1-\alpha)} + \frac{u}{2} - C, & u < 0 \end{cases} \tag{2.10}$$

and the subset of points in $\mathbb{R}^2$ lying between the graphs of B_1 and B_2:

$$S = \{(u,v) : v \leq B_1(u), v \geq B_2(u)\}. \tag{2.11}$$

Finally, suppose that γ in the linear rule $F_\gamma(u,v) = u + \gamma v$ is set to a value in the range,

$$\frac{2\big(2C(1-\alpha^2)\big)^{1/2} - (1+\alpha)\}}{\big(2C(1-\alpha^2)\big)^{1/2} + 2\alpha C} \leq \gamma \leq \frac{4(1+\alpha)((1+\alpha) + 2)}{8C(1+\alpha) - (1+\alpha) - 4}.$$

If (u_{n+1}, v_{n+1}) is obtained from $(u_n, v_n) \in S$ via the second-order recursion (2.5) with input f_n satisfying $|f_n| \leq \alpha$, quantizer Q_{tri}^τ, $\tau \leq 1/2$, and linear rule $F_\gamma(u_n, v_n)$, then $(u_{n+1}, v_{n+1}) \in S$.

A qualitative depiction of the invariant region S is shown in Figure 2.1 below, as generated by admissible parameter values $\alpha = .9$, $C = 40$, and $\gamma = .1$.

2.0.8 The revised scheme: Quiet $\Sigma\Delta$ quantization

Henceforth, we will assume that the oversampling factor λ is fixed, and that the damping factor $\rho < 1$ in the finite-memory scheme (2.5) satisfies the requirements of
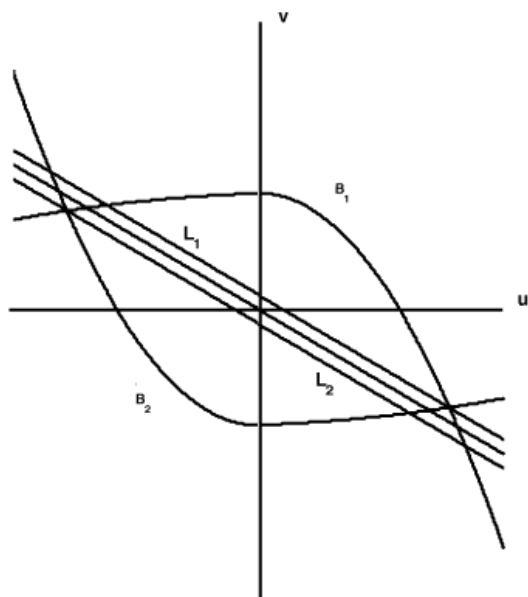
Figure 2.1: The functions B_1 and B_2 from Theorem 2.0.15 are pictured along with the lines L_1 and L_2 consisting of the points (u, v) such that $F_\gamma(u, v) = .5$ and $F_\gamma(u, v) = -.5$, respectively. This figure was generated using admissible values $(\alpha, C, \gamma) = (.9, 40, .1)$

.

Lemma 2.0.14 for the finite-memory scheme to be second-order. We further assume that the ℓ_∞ bound $|f_n| \leq \alpha < 1$ is fixed, and that γ is in the admissible range, as specified by Theorem 2.0.15, for the finite-memory $\Sigma\Delta$ scheme (2.5) to be stable.

Unfortunately, the finite-memory scheme (2.5) implemented with tri-level quantizer $Q_{tri}^{.5}$ is not quiet for ρ sufficiently close to 1, as a range of initial conditions (u_0, v_0) at the onset of zero input still lead to oscillatory behavior of the output. In Figure 2.2, we indicate whether or not the iterates (u_n, v_n) generated by the finite-memory scheme with zero input $x_n = 0$ remain oscillatory, for a set of perturbations $\rho \in [.96, 1]$ and initial conditions $(u_0, 0) \in [-2, 0] \times 0$.

Nevertheless, we show in the following sections that a simple modification to the finite-memory scheme (2.5) will ensure quietness, in the sense of Definition 2.0.12. The key idea is to apply a damping factor ρ as in the finite second order scheme (2.5), but not at *every* iteration; only at iterations n for which $u_n + \gamma v_n \geq 0$ (or the symmetric, when $u_n + \gamma v_n \leq 0$). The situation $u_n + \gamma v_n \geq 0$ must keep reccurring if the system is to remain stable, and the application of an asymmetric damping *forces* the state variables u_n, and in turn the variables v_n, to zero once $f_n = 0$.

More precisely, we will consider the following *asymmetrically damped* modification to (2.5), which, starting from initial (u_0, v_0), can be implemented according to the
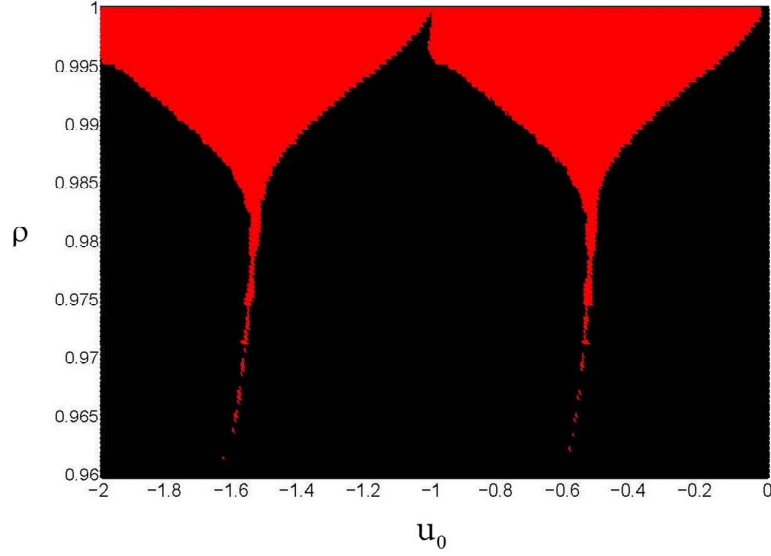
Figure 2.2: We indicate whether or not the iterates (u_n, v_n) converge to zero, as generated by the second order finite-memory scheme (2.5) with zero input $x_n = 0$, for several values of damping factor $\rho \in [.96, 1]$ and initial conditions $(u_0, 0) \in [-2, 0] \times 0$. In particular, $(u_n, , v_n) \to 0$ can only occur over the black region, as a gray mark indicates that after one million iterations $n \geq 1000000$, the quantization output q_n is nonzero $|q_n| = 1$ at regular time interval not exceeding 100 iterations, for one million additional steps $n \in [1000000, 2000000]$. The finite-memory scheme (2.5) was implemented with parameter $\rho = .2$ and tri-level quantizer $Q_{tri}^{1/3}$.

.

recursion

$$
\begin{aligned}
(b_n, q_n) &= Q_4^\tau(u_{n-1}/\gamma + v_{n-1}), \\
u_n &= u_{n-1} + q_n(\rho - 1)u_{n-1} - b_n + f_n^\lambda, \\
v_n &= v_{n-1} + q_n(\rho - 1)v_{n-1} + u_n,
\end{aligned}
\tag{2.12}
$$

with 4-level quantizer,

$$
Q_4^\tau(u) = \begin{cases}
(-1, 0), & u \leq -1/2, \\
(0, 0), & -1/2 < u \leq 0, \\
(0, 1), & 0 < u \leq 1/(2\tau), \\
(1, 1), & u > 1/(2\tau).
\end{cases}
\tag{2.13}
$$

Although we will take $\tau = \rho$ for the theoretical analysis of the next section, simulation results suggest that quietness persists for a much more general class of 4-level quantizers that includes the symmetric quantizer Q_4^1.

That the asymmetric scheme (2.12) maintains second order accuracy for $\rho \geq \rho^\lambda = 1 - \frac{1}{\lambda}$ is an immediate consequence of Proposition 2.0.14. Furthermore, the asymmetric scheme is stable and inherits the positively invariant sets S_γ constructed in Proposition 2.0.15. It remains to prove that (2.12) is quiet. Our accomplishments are summarized in the following theorem.

Theorem 2.0.16 (Main theorem). *The asymmetrically-damped second-order $\Sigma\Delta$ scheme (2.12) is quiet when implemented with 4-level quantizer Q_4^ρ. That is, if the*

input (f_n) to (2.12) becomes identically equal to zero after some time $n \geq N$, then the quantization output $(b_n, q_n) \equiv (0, 1)$ becomes constant after a finite number of additional iterations $n \geq N + k$.

We prove Theorem 3.5.1 by showing that $(0, 0)$ is a global attracting fixed point of the zero-input map governing $(u_n, v_n) \rightarrow (u_{n+1}, v_{n+1})$ in the asymmetric scheme (2.12). As such, we need to develop a better understanding for this dynamical system.

Observation 2.0.17. *The discrete map $(u_n, v_n) \rightarrow (u_{n+1}, v_{n+1})$ produced by the asymmetrically-damped $\Sigma\Delta$ scheme (2.12) with zero input $f_n^\lambda \equiv 0$ and 4-level quantizer Q_4^ρ corresponds to the orbit of a piecewise-affine dynamical system,*

$$(u_n, v_n) = M_{(\gamma,\rho)}(u_{n-1}, v_{n-1}) = A_\gamma \circ D_{(\gamma,\rho)}(u_{n-1}, v_{n-1}), \tag{2.14}$$

where

$$D_{(\gamma,\rho)}(u, v) = \begin{cases} (\rho u, \rho v), & u/\gamma + v > 0, \\ (u, v), & else, \end{cases} \tag{2.15}$$

and

$$A_\gamma(u, v) = \begin{cases} (u - 1, u + v - 1), & u/\gamma + v > 1/2, \\ (u, u + v), & |u/\gamma + v| \leq 1/2, \\ (u + 1, u + v + 1), & u/\gamma + v < -1/2. \end{cases}$$

The origin $(0, 0)$ is clearly a fixed point of the map $M_{(\gamma,\rho)}$. We will show that $(0, 0)$ is a global attracting fixed point, or that $(u_n, v_n) = M_{(\gamma,\rho)}^n(u_0, v_0) \rightarrow (0, 0)$ for any initial condition (u_0, v_0), in two steps:

1. In Section 2.0.9, we construct a positively invariant set T for the map $M_{(\gamma,\rho)}$ having the property that all orbits initialized in T converge to the origin.

2. In Section 2.0.10, we show that T is also a *trapping set* for $M_{(\gamma,\rho)}$, so that all orbits eventually become trapped in T.

It is not clear that the proof need necessarily be split into two steps; however, numerical results such as those in Figure 2.5 suggest a marked change in the behavior of sequences $(\mathbf{x}, M_{(\gamma,\rho)}\mathbf{x}, M^2_{(\gamma,\rho)}\mathbf{x}, ...)$ upon entering the invariant set T.

2.0.9 An invariant set T and convergence to the origin

We begin by constructing a positively invariant set for the map $M_{(\gamma,\rho)}$. We refer the reader to Figure 2.3 for reference.

Proposition 2.0.18. *The union $T = T^+ \cup T^-$ of the affine regions*

$$
\begin{aligned}
T^+ &= \{(u,v): 0 < u < 1, \ -1/2 \le u/\gamma + v \le (1/2 + 1/\gamma)\}, \\
T^- &= \{(u,v): -1 < u \le 0, \ -(1/2 + 1/\gamma) \le u/\gamma + v \le 1/2\}
\end{aligned}
$$

is a positively invariant set for the map $M_{(\gamma,\rho)}$. Moreover, the bit sequence b_n as in (2.12) associated to a sequence $M^n_{(\gamma,\rho)}(u_0,v_0) = (u_n,v_n)$ contained in T has the following alternating structure:

- *If $b_n = 1$ then $b_{n+1} \in \{-1,0\}$,*

- *If $b_n = -1$ then $b_{n+1} \in \{0,1\}$.*
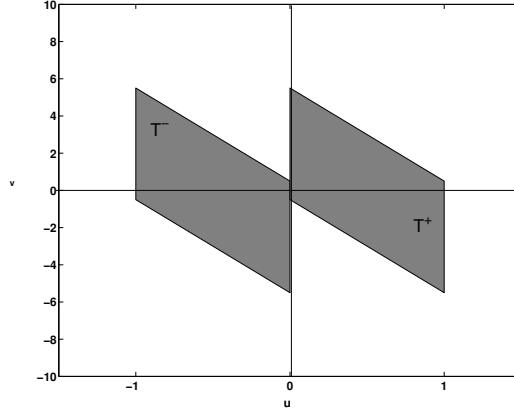
(a) The positively invariant set T

Figure 2.3: The positively invariant set $T = T^+ \cup T^-$ for the map $M_{(\gamma,\rho)}$ with parameter $\gamma = .2$.

Finally, $b_n = 1$ if and only if $(u_n, v_n) \in T^+$ and $(u_{n+1}, v_{n+1}) \in T^-$, while $b_n = -1$ if and only if $(u_n, v_n) \in T^-$ and $(u_{n+1}, v_{n+1}) \in T^+$.

Proof. First, as $T = T^+ \cap T^-$ is the union of two convex sets, each containing the origin, $\mathbf{x} \in T$ implies $\lambda \mathbf{x} \in T$ for $\lambda \in [0, 1]$. As such, if T is positively invariant for the map A_γ, then T must also be positively invariant for the full map $M_{(\gamma,\rho)} = A_\gamma \circ D_{(\gamma,\rho)}$. In order to show that T is positively invariant for A_γ, or that $(u, v) \in T$ implies $A_\gamma(u, v) \in T$, we can assume without loss that that $(u, v) \in T^+$ by symmetry of T and A_γ. Note that the bit b in (2.12) associated to a point $(u, v) \in T^+$ in the asymmetric scheme (2.12) is restricted to $b \in \{0, 1\}$. We split the proof into two cases, according to whether $b = 0$ or $b = 1$.

1. **Case 1**: $b = 1$, or $\gamma/2 < u + \gamma v \leq \gamma/2 + 1$: In this case, $(u', v') = A_\gamma(u, v) = (u - 1, v + u - 1)$. Since $-1 < u' < 0$, $(u', v') \in T$ if $(u', v') \in T^-$, or if

71

$(u', v') \in T^-$ if $-(\gamma/2 + 1) \le u' + \gamma v' \le \gamma/2$. We can expand $u' + \gamma v'$ as

$$u' + \gamma v' = u + \gamma v - (\gamma + 1) + \gamma u,$$

and so indeed

$$u' + \gamma v' \le (\gamma/2 + 1) - (\gamma + 1) + \gamma \le \gamma/2,$$

while also

$$u' + \gamma v' \ge \gamma/2 - (\gamma + 1) + \gamma u \ge -(\gamma/2 + 1).$$

Thus, $(u', v') \in T^-$ and has associated bit $b' \in \{-1, 0\}$.

2. **Case 2:** $b = 0$, or $-\gamma/2 < u + \gamma v \le \gamma/2$: Within this region, $(u', v') = A_\gamma(u, v) = (u, v + u)$. Since $0 < u' \le 1$, $(u', v') \in T$ if $(u', v') \in T^+$, or if $-\gamma/2 \le u' + \gamma v' \le \gamma/2 + 1$. Proceeding as before,

$$u' + \gamma v' = u + \gamma v + \gamma u \le \gamma/2 + \gamma \le \gamma/2 + 1,$$

while at the same time

$$u' + \gamma v' > -\gamma/2 + \gamma u > -\gamma/2.$$

Thus, $(u', v') \in T^+$ and has associated bit $b' \in \{0, 1\}$.

The proposition follows by incorporating the symmetric result for initial conditions $(u, v) \in T^-$. $\qquad \square$

The alternating pattern of the bit sequence b_n associated to iterates $(u_n, v_n) \in T$ will be key in the following lemma. In particular, this constraint on the bit sequence, in combination with the imbalance induced by damping $(u_{n+1}, v_{n+1}) = (\rho u_n, \rho v_n)$ only in the positive half-space $S^+ = \{(u, v) : u + \gamma v \geq 0\}$, forces the u_n, and in turn the v_n, to zero.

Lemma 2.0.19. *Suppose that $(u_0, v_0) \in T$. Then the subsequence (u_{n_j}) from $(u_n, v_n) = M^n_{(\gamma, \rho)}(u_0, v_0)$ consisting of indices n_j for which $u_{n_j} \in T^+$ satisfies $u_{n_j} \to 0$ as $n \to \infty$.*

Proof. First note that the event $(u_n, v_n) \in T^+$ must keep recurring, for if not, or if $(u_n, v_n) \notin T^+$ for all $n \geq 0$, then $b_n = 0$ for all n according to Proposition 2.0.18, and

$$v_n = v_0 + \sum_{j=0}^{n} u_j = v_0 + (n+1)u_0$$

diverges. The index set $\mathcal{I}^+ = \{n_j : u_{n_j} \in T^+\}$ is then an infinite subset of the natural numbers, so that (u_{n_j}) represents an infinite subsequence of (u_n). As $u_{n+1} = \rho u_n - 1$ in passing from T^+ to T^-, $u_{n+1} = u_n$ while both (u_n, v_n) and (u_{n+1}, v_{n+1}) remain in T^-, and $u_{n+1} = u_n + 1$ in passing back from T^- to T^+, it follows that $u_{n_{j+1}} = \rho u_{n_j}$ contracts along the index set $\mathcal{I}^+$. The subsequence $u_{n_j} = \rho^j u_{n_0}$ thus converges to zero as $n_j \to \infty$. $\qquad\square$

Note that convergence of the subsequence (u_{n_j}) in Lemma 2.0.19 does not guarantee convergence of the *full* sequence (u_n), as the residual sequence (u_{n_k}) over the index set $\mathcal{I}^- = \mathbb{N} \setminus \mathcal{I}^+$ forms a subsequence satisfying $u_{n_k} \to -1$ as $k \to \infty$, if $\mathcal{I}^-$ is an infinite set. However, the following proposition ensures that this situation cannot occur.

Proposition 2.0.20. *If $(u_0, v_0) \in T$, then the full state sequence $(u_n, v_n) = M^n_{(\gamma, \rho)}(u_0, v_0)$ satisfies $(u_n, v_n) \to (0^+, 0^+)$ as $n \to \infty$.*

Proof. Using the aforementioned results, we assume without loss that $(u_0, v_0) \in T^+$, and that $u_0 \leq \epsilon$ for arbitrarily small $\epsilon > 0$.

1. Consider the situation where $(u_n, v_n) \in T^+$ for all $n \geq 0$. In this case, $b_n = 0$ for all $n \geq 0$, $u_n \to 0^+$, and v_n is defined recursively according to

$$
\begin{aligned}
v_{n+1} &= \rho v_n + \rho u_{n-1} \\
&= \rho^2 v_{n-1} + \rho^2 u_{n-2} + \rho u_{n-1} \\
&= \ldots \\
&= \rho^n v_0 + \sum_{j=0}^{n-1} \rho^{n-j} u_j \\
&= \rho^n v_0 + \rho^n \epsilon, \quad\quad\quad (2.16)
\end{aligned}
$$

   where we use in the last equality that $u_j = \rho^j u_0$. In this case, then, $v_n \to 0$.

2. Suppose that $\rho(u_0 + \gamma v_0) \leq \gamma/2$, in which case $b_1 = 0, u_1 = \rho u_0$, and $v_1 = \rho(v_0 + u_0)$, yielding

$$
\begin{aligned}
\rho(u_1 + \gamma v_1) &= \rho^2(u_0 + \gamma v_0) + \rho^2 \gamma u_0 \\
&\leq \rho(\gamma/2) + \rho^2 \gamma \epsilon, \\
&\leq \gamma/2, \quad\quad\quad (2.17)
\end{aligned}
$$

   the last inequality requiring that $u_0 \leq \epsilon \leq (1 - \rho)/(2\rho^2)$, which is satisfied as

74

ϵ was arbitrary. This case, then, reduces to the first case.

3. It remains only to consider sequences (u_n, v_n) for which $(u_n, v_n) \in T^+$ implies $b_n = 1$, or $\rho(u_n + \gamma v_n) > \gamma/2$. In terms of Figure 2.3, such sequences avoid the lower parallel section of the set T^+. The trajectory of such (u_n, v_n) under the map $M_{(\gamma, \rho)}$ can then be described as follows:

(a) If $(u_n, v_n) \in T^+$, then

$$b_n = 1,$$

$$u_{n+1} = \rho u_n - 1,$$

$$v_{n+1} = \rho v_n + \rho u_n - 1 \leq \rho v_n + \rho \epsilon - 1$$

Else if $(u_n, v_n) \in T^-$,

(b) If $-1/2 \leq u_n/\gamma + v_n \leq 1/2$, then

$$b_n = 0,$$

$$u_{n+1} = u_n \leq \epsilon - 1,$$

$$v_{n+1} = v_n + u_n \leq v_n + \epsilon - 1$$

(c) If $u_n/\gamma + v_n \leq -1/2$, then

$$b_n = -1,$$

$$u_{n+1} = u_n + 1 \leq \epsilon,$$

$$v_{n+1} = v_n + u_n + 1 \leq v_n + \epsilon$$

75

Since (c) cannot occur in successive iterations according to the alternating nature of b_n, we arrive at the period-2 inequality

$$v_{n+2} \leq v_n + 2\epsilon - 1, \tag{2.18}$$

indicating that the iterates v_n diverge. This case, in conclusion, cannot occur.

$\square$

2.0.10 The invariant set T as a global trapping set for $M_{(\gamma,\rho)}$

We prove in this section that the positively invariant set T is also a global *trapping set* for $M_{(\gamma,\rho)}$; this result in combination with the results of the last section guarantee that the origin is a global fixed point for $M_{(\gamma,\rho)}$.

To be precise,

Definition 2.0.21. *A global trapping set of a discrete mapping $M : \mathbb{R}^2 \to \mathbb{R}^2$ is any set $S \subset \mathbb{R}^2$ such that*

1. *$M(S) \subset S$ (that is, S is a positively invariant set for M),*

2. *for any $\mathbf{x} \in \mathbb{R}^2$, there exists $n \geq 0$ such that $M^n \mathbf{x} \in S$.*

The construction of global trapping sets for the *full* second-order $\Sigma\Delta$ scheme was recently developed by Sidong [80]. We will follow the approach of that work and show that T is a global trapping set for the map $M_{(\gamma,\rho)}$ using a *Lyapunov function* argument. In the context of discrete dynamical systems, a Lyapunov function refers

loosely to a nonnegative convex energy functional $h : \mathbb{R}^2 \to \mathbb{R}$ that contracts along the action of the discrete map, $h(M\mathbf{x}) \leq (1 - \delta)h(\mathbf{x})$. If h contracts as such for all $\mathbf{x} \in \mathbb{R}^2$, and $h(\mathbf{x}) = 0$ only if $\mathbf{x} = \mathbf{0}$, then all orbits $(u_n, v_n) = M^n(u_0, v_0)$ must converge to the global fixed point $(0, 0)$. More generally, if h decreases along orbits $h(M^n\mathbf{x}) \leq h(\mathbf{x}) - \delta$ while the iterates $M^n\mathbf{x}$ remain outside an invariant set $S \in \mathbb{R}^2$, then S can be shown to be a globally attracting set, as follows:

Lemma 2.0.22. *Let S be a positively invariant set for the discrete map M. Suppose there exists a nonnegative function $h : \mathbb{R}^2 \to \mathbb{R}^+$ and a parameter $\delta > 0$ for which, if $\mathbf{x}$ is not in S, then either $M^k\mathbf{x} \in S$ or $h(M^k\mathbf{x}) - h(\mathbf{x}) \leq -\delta$ after some finite time k. It follows that S is a global trapping set for M.*

Proof. Suppose $\mathbf{x} \notin S$ is such that $M^k\mathbf{x} \notin S$ for all $k \geq 0$, and denote $c = h(\mathbf{x})$. From the stated hypotheses, $h(M^{k_1}\mathbf{x}) \leq c - \delta$ after some finite time k_1, and, by induction, $h(M^{k_n}\mathbf{x}) \leq c - n\delta$ after a finite time k_n for any positive integer n. But then eventually $h(M^k\mathbf{x}) \leq 0$, which is impossible since $h \geq 0$ by assumption. $\square$

Lemma (2.0.22) has the following implication for the map $M_{(\gamma, \rho)} = A_\gamma \circ D_{(\gamma, \rho)}$ and invariant set T under consideration.

Proposition 2.0.23. *Consider the positively invariant set T as constructed in Proposition (2.0.18). T is a global trapping set for $M_{(\gamma, \rho)}$ if there exists a nonnegative and convex function $h : \mathbb{R}^2 \to \mathbb{R}^+$ and a parameter $\epsilon > 0$ for which the following sets are*

contained in T:

$$\Delta_h \quad : \quad \{(u,v) \in \mathbb{R}^2 : h(u,v) \le h(A_\gamma(u,v))\},$$

$$\Omega_h^\epsilon \quad : \quad \{(u,v) \in \mathbb{R}^2 : h(u,v) \le \epsilon\} \tag{2.19}$$

Proof. We verify that the conditions for Lemma (2.0.22) hold for invariant set T and parameter $\delta = \epsilon(1 - \rho) > 0$. The proof is split into two cases, according to whether or not $\mathbf{x} \in S^+ = \{(u,v) \in \mathbb{R}^2 : u + \gamma v \ge 0\}$ or $S^- = \mathbb{R}^2 \setminus S^+$.

1. **Case 1:** Suppose first that $\mathbf{x} \in S^+$. On this half-space, $D_{(\gamma,\rho)} : \mathbf{x} \to \rho\mathbf{x}$ acts as a contraction. If $\mathbf{x} \notin T$, but $D_{(\gamma,\rho)}\mathbf{x} \in T$, then $M_{(\gamma,\rho)}\mathbf{x} = A_\gamma \circ D_{(\gamma,\rho)}\mathbf{x} \in T$ by positive invariance of T for A_γ. If alternatively $D_{(\gamma,\rho)}\mathbf{x} \notin T$, then

$$
\begin{aligned}
h(M_{(\gamma,\rho)}\mathbf{x}) - h(\mathbf{x}) \quad &= \quad h(A_\gamma(D_{(\gamma,\rho)}\mathbf{x})) - h(\mathbf{x}) \\
&\le \quad h(D_{(\gamma,\rho)}\mathbf{x}) - h(\mathbf{x}), \text{ as } \Delta_h \in T \\
&= \quad h(\rho\mathbf{x}) - h(\mathbf{x}) \\
&\le \quad \rho h(\mathbf{x}) - h(\mathbf{x}), \text{ using convexity of } h \text{ and } h(\mathbf{0}) = 0 \\
&\le \quad -\epsilon(1 - \rho), \text{ since } \Omega_h^\epsilon \in T \\
&= \quad -\delta.
\end{aligned}
$$

2. **Case 2**: Suppose now that $M_{(\gamma,\rho)}^k \mathbf{x} \in (S^+)^c \cap T^c$ for all $k \ge 0$. Under this assumption, the quantization output is restricted to $b_j \in \{-1, 0\}$ for all j, so

78

that $u_j = u_{j-1} - b_j$ forms a monotonically nondecreasing sequence. If $u_j < 0$ for all j, then $|v_j| = |v_0 + \sum_{k=0}^{j} u_k|$ diverges, which is impossible since (u_j, v_j) belongs to a bounded set. If alternatively $u_j > 0$ at some index j, then $u_n > 0$ for all $n \geq j$ by monotonicity and, still, $|v_j| \to \infty$. The case $u_j = 0$ for all $j \geq n$ cannot happen since $(u_n, v_n) \notin T$ by assumption. It follows that this case is impossible.

We can conclude that after a finite number of iterations k, either $M_{(\gamma,\rho)}^k \in T$ or $M_{(\gamma,\rho)}^k \in S^+$. The result follows by application of Lemma 2.0.22. $\quad\square$

Following the approach in [80], we consider the following Lyapunov function for the map A_γ:

$$h(u, v) = u^2 + |2v - u|. \tag{2.20}$$

Note that h is a convex function on $\mathbb{R}^2$. Letting $h^+(u, v) = u^2 + 2v - u$ and $h^-(u, v) = u^2 - 2v + u$, it is easily verified that $h(u, v) = \max\{h^+(u, v), h^-(u, v)\}$. The motivation for this choice of Lyapunov function is that h^+ and h^- are the unique functional solutions satisfying the relations

$$
\begin{aligned}
h^+(A_\gamma(u, v))\big|_{u+\gamma v>\gamma/2} = h^+(u - 1, u + v - 1) &= h^+(u, v), \\
h^-(A_\gamma(u, v))\big|_{u+\gamma v<-\gamma/2} = h^-(u + 1, u + v + 1) &= h^-(u, v).
\end{aligned}
\tag{2.21}
$$

It is straightforward that the positively invariant set T contains the set $\Omega_h^\epsilon = \{(u, v) : h(u, v) \leq \epsilon\}$ for sufficiently small ϵ. In order to apply Proposition (2.0.23), it remains only to verify that T also contains the set $\Delta_h = \{(u, v) \in \mathbb{R}^2 : h(u, v) \leq h(A_\gamma(u, v))\}$.

Proposition 2.0.24. *Consider the map A_γ as defined in (2.0.17), the convex function $h(u, v) = u^2 + |2v - u|$, and the set $R = R_1 \cup R_2$, with*

$$R_1 = \{(u, v) \in \mathbb{R}^2 : u + \gamma v \geq 0, 2v + u \leq 1, u \leq 1/2\}, \text{ and}$$

$$R_2 = \{(u, v) \in \mathbb{R}^2 : u + \gamma v < 0, 2v + u \geq -1, u \geq -1/2\}. \qquad (2.22)$$

The set R is contained in the positively invariant set T (as shown in Figure 2.4). Moreover, R contains the set $\Delta_h = \{(u, v) \in \mathbb{R}^2 : h(u, v) \leq h\big(A_\gamma(u, v)\big)\}$.

We defer the proof of Proposition 2.0.24, which amounts to a straightforward case by case analysis, to Appendix 1.

The main result of this chapter, Theorem 3.5.1, follows from Proposition 2.0.23 and Lemma 2.0.20.

Proof of Theorem 3.5.1. By Proposition (2.0.23), T is a global trapping set for $M_{(\gamma,\rho)}$, the discrete map governing $(u_n, v_n) \to (u_{n+1}, v_{n+1})$ in the asymmetrically-damped scheme (2.12) with 4-level quantizer Q_4^ρ and zero input $f_n \equiv 0$. In other words, for any initial condition (u_0, v_0), the iterates $(u_n, v_n) = M_{(\gamma,\rho)}^n(u_0, v_0)$ become trapped in the set T after a finite number of iterations. From Lemma 2.0.20, it follows that $(u_n, v_n) \to (0^+, 0^+)$; once the iterates $(u_n, v_n) \in T^+ \cap \{0 \leq u + \gamma v \leq \gamma/2\}$ become trapped in the positive quadrant, the bit output $(b_n, q_n) = (0, 1)$ is constant. $\qquad \square$
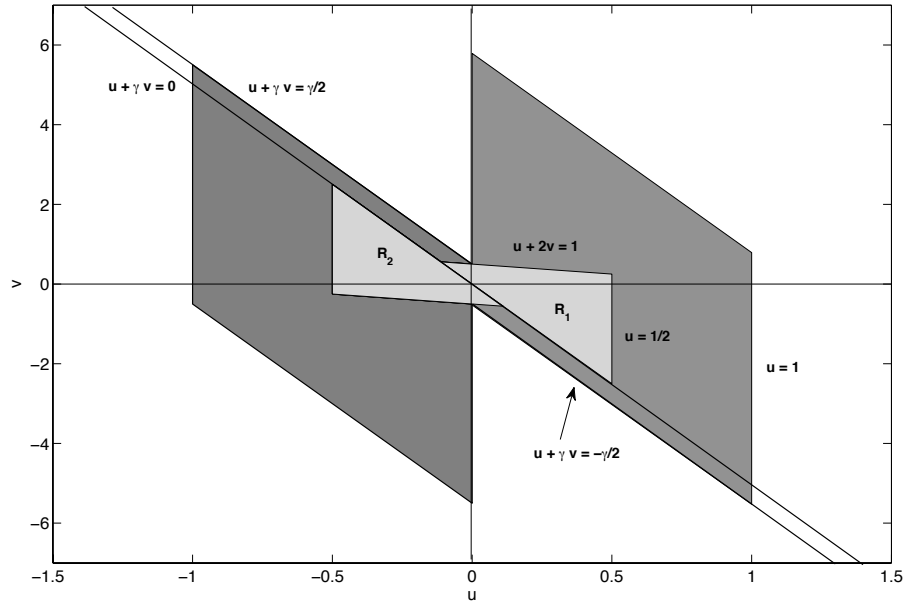
Figure 2.4: The positively invariant set T is the union of two parallelograms, as shown in dark gray for parameter $\gamma = .2$. As the region Δ_h, on which the map A_γ may satisfy $A_\gamma(h(u,v)) - A_\gamma(u,v) > 0$, is contained in R (light gray triangles) which is in turn contained in T, the set T is a globally attracting set.
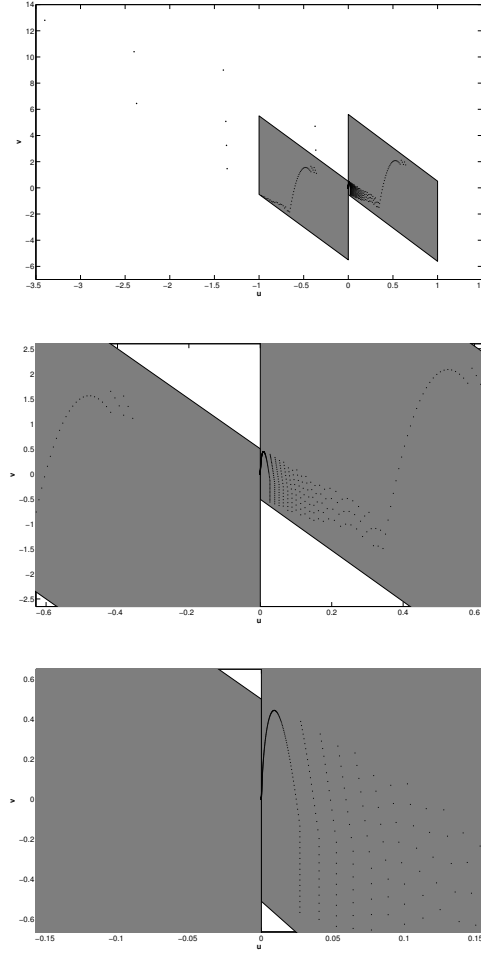
Figure 2.5: Different magnifications of an orbit of the map $M_{(\gamma,\rho)}$ for $\rho = .98$ and $\gamma = .2$. The initial point $(u_0, v_0) = (-3.4, 12.7)$ can be seen in the top image; this point and the first few iterations are outside the trapping set T (dark gray parallelograms). Once trapped in T, the iterates (u_n, v_n) converge to the globally attracting fixed point $(0, 0)$.

2.1   Future Directions

We have introduced a 2-bit second order $\Sigma\Delta$ scheme that is guaranteed to be *quiet*: periodicities in the bit output are eliminated when the input vanishes. It remains to analyze the stability of such quietness in the face of unavoidable component imprecisions. Although a full analysis is beyond the scope of the current presentation, we outline a few key issues below.

1. **Randomness helps.** Any introduction of randomness to a quantization scheme will only *increase* the aperiodicity of the bit output; this includes zero-mean additive noise on the samples $x_n = f_n + \epsilon_n$ and unbiased 'flakiness' in the quantizer element. Quietness for the asymmetrically-perturbed scheme (2.12) is threatened more from *biased* imprecisions, such as those resulting from an offset in the 4-level quantizer, $\tilde{Q}_4^\rho(u) = Q_4^\rho(u + \delta)$. Numerical simulations suggest that quietness of the asymmetric scheme nevertheless persists in the face of such offsets: the state sequence (u_n, v_n) still converges at the onset of zero input, but the iterates v_n approach a nonzero limit $L(\delta)$ that grows with the magnitude of the shift; if $\delta > 0$ is small enough, then quietness is still attained. More generally, quietness of the asymetrically-damped scheme appears to persist for a much more general class of quantizers than the the particular 4-level quantizer Q_4^ρ that was needed for the theoretical analysis of the previous section; indeed, indistinguishable results are observed by implementing (2.12) with the far simpler symmetric 4-level quantizer,

$$Q_4(u) = \begin{cases} (-1,0), & u \leq -1/2, \\ (0,0), & -1/2 < u \leq 0, \\ (0,1), & 0 < u \leq 1/2, \\ (1,1), & u > 1/2. \end{cases} \tag{2.23}$$

2. **Other robustness issues.**

   (a) **Imprecisions in ρ:** The damping factor ρ does not have to be constant from iteration to iteration nor known precisely; all of the analysis of the previous section holds for $\rho_n > \rho^\lambda$ varying at each iteration, as long as the sequence remains bounded from below as to maintain second-order accuracy of the scheme (2.12), and is not chosen adversarialy to converge $\rho_n \to 1$.

   (b) **Imprecisions in γ:** We have until now assumed that the parameter γ is fixed throughout the iterations. However, numerical evidence suggests that quietness does not require that γ be fixed in the expression $u_n + \gamma v_n$, as long as γ is within the range of stability, as stated in Theorem (2.0.15).

   (c) **Leaky integrators.** More realistically, one should incorporate integrator leakage into the proposed model (2.12), analogous to the modification studied for the golden ratio encoder in the last chapter. That is, the asymmetric scheme, being second order, requires two delay elements in its implementation to hold each of the states u_n and v_n over one iteration. After one clock time, the stored input in the first delay is reduced to λ_1 times its original value, while the stored input in the second delay is

replaced by λ_2 times its original value; we can incorporate such leakage into the asymmetric model (2.12) as follows:

$$
\begin{aligned}
(b_n, q_n) &= Q^4(\lambda_1 u_{n-1}/\gamma + \lambda_2 v_{n-1}), \\
u_n &= \lambda_1\big(u_{n-1} + q_n(\rho_1 - 1)u_{n-1}\big) - b_n + f_n^\lambda, \\
v_n &= \lambda_2\big(v_{n-1} + q_n(\rho_2 - 1)v_{n-1}\big) + u_n.
\end{aligned}
\tag{2.24}
$$

Above, $(\lambda_1, \lambda_2) \in [.95, 1]^2$ in most circuits on interest; the specific leakage factors within this window are generally unknown and may vary slowly in time. When $\lambda_1 = \lambda_2$, the positively invariant set T is still a trapping set for the revised scheme; however, once in this set, it is not immediately clear how to modify the analysis of section 2.0.9 to prove that $(u_n, v_n) \to (0, 0)$. Numerical studies indicate that quietness persists in the face of such leakage, requiring only that the perturbations remain asymmetric, or that each of $\lambda_1 \rho_n$ and $\lambda_2 \rho_n$ is strictly smaller than either of λ_1 and λ_2.

3. Hybrid Chaotic-Quiet $\Sigma\Delta$ Quantization

While the asymmetry of the perturbation $1 - \rho_n$ in (2.12) is key for inducing quietness, it is not clear that this perturbation need be nonnegative. So-called *chaotic $\Sigma\Delta$ quantization* [51] has been proposed as a means to eliminate idle tones in $\Sigma\Delta$ quantization, not just at the onset of zero input, but at the onset of any constant, or DC input sequence. In the context of 1-bit, second-order $\Sigma\Delta$ quantization, chaotic $\Sigma\Delta$ quantization corresponds to *expanding*, rather

than contracting, by a small factor $(1 + \epsilon) > 0$ at each iteration:

$$
\begin{aligned}
b_n &= Q(u_{n-1} + \gamma v_{n-1}), \\
u_n &= (1 + \epsilon)u_{n-1} + f_n^\lambda - b_n, \\
v_n &= (1 + \epsilon)v_{n-1} + u_n.
\end{aligned}
\tag{2.25}
$$

This modification is termed chaotic because it has the effect of generating aperiodic output at the onset of DC input; however, a complete stability analysis of the scheme (2.25) remains an open problem. Numerical studies, such as those in Figure 2.1, indicate that the chaotic scheme outperforms the asymmetrically-damped scheme (2.12) in generating aperiodic output for general DC input, although the asymmetric scheme shows a marked improvement over the fully damped tri-level scheme. While the chaotic scheme (2.25) is not quiet in the sense of Definition (2.0.12), we expect that quietness can be achieved by introducing asymmetry into the model (2.25). Specifically, the hybrid scheme,

$$
\begin{aligned}
(b_n, q_n) &= Q_4(u_{n-1}/\gamma + v_{n-1}), \\
u_n &= u_{n-1} + q_n \epsilon u_{n-1} - b_n + f_n^\lambda, \\
v_n &= v_{n-1} + q_n \epsilon v_{n-1} + u_n,
\end{aligned}
\tag{2.26}
$$

with q_n taking values in $\{-1, 1\}$ so that either damping $(\rho_n = 1 - \epsilon)$ or expansion $(\rho_n = 1 + \epsilon)$ is applied at each iteration, appears to inherit the aperiodic orbit structure of its parent chaotic map (2.25), and also quietness induced by its

asymmetry. We hope to study the stability and aperiodicity for the chaotic maps (2.25), (2.26) in the future.
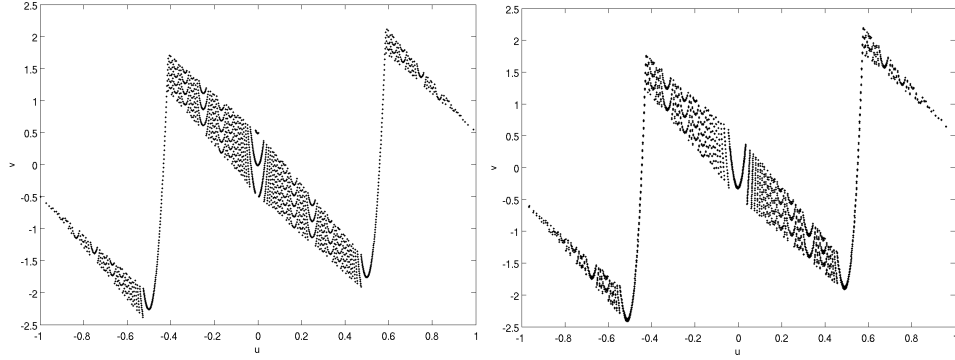
(a) Original and damped second order $\Sigma\Delta$ scheme (2.5) with $\rho = 1$ and $\rho = .995$, respectively
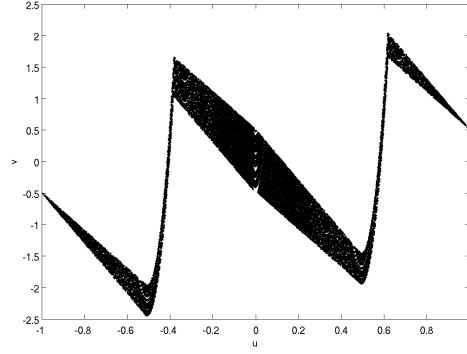
(b) Asymmetrically damped $\Sigma\Delta$ scheme with $\rho = .995$.
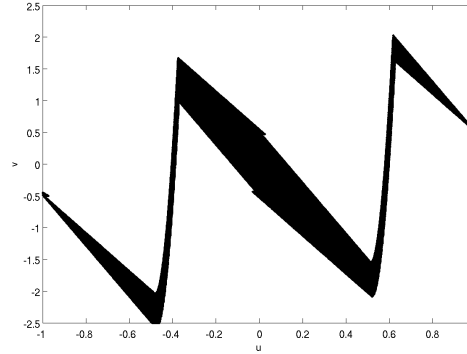
(c) Chaotic $\Sigma\Delta$ scheme with $\rho = 1.01$.

Figure 2.6: One million iterates (u_n, v_n) from a single orbit under (a) the original damped second order scheme (2.5) with damping factor $\rho = 1$ and $\rho = .995$, respectively, (b) the asymmetrically-damped scheme (2.12), with asymmetric damping factor $\rho = .995$, and (c) the chaotic map (2.25), with expansion factor $\rho = 1.01$. Constant input $f_n = -.001$ was used for all cases, along with parameter $\gamma = .2$ and either tri-level quantizer (a) or 4-level quantizer $(b), (c)$. The asymmetrically-damped orbit (b) is seen to fill up more space in the plane than the fully damped orbits (a), but not as densely as the chaotic orbit (c), which appears to be aperiodic.

2.2 Appendix : proof of Proposition 2.0.24

Let $S^+ = \{(u,v) \in \mathbb{R}^2 : u + \gamma v \geq 0\}$ and let $S^- = \{(u,v) \in \mathbb{R}^2 : u + \gamma v < 0\} = \mathbb{R}^2 \setminus S^+$. Suppose that $(u,v) \in S^+ \setminus R_1$ is such that $(u',v') = A_\gamma(u,v) = (u-1, u+v-1)$. Our first aim is to show that $h(M_1(u,v)) \leq h(u,v)$ in this situation.

1. **Case 1:** If $2v - u \geq 0$, then $h(u,v) = u^2 + 2v - u$, while $h(u',v') = u^2 - 2u + 1 + |u + 2v - 1|$,so

$$h(u',v') \leq h(u,v) \iff |u + 2v - 1| \leq u + 2v - 1 \iff u + 2v \geq 1.$$

   But since $(u,v) \in S^+ \setminus R_1$, we know that $u > 1/2$, so $u + 2v \geq 2u > 1$, and $h(u',v') \leq h(u,v)$ holds in this case.

2. **Case 2:** If $2v - u \leq 0$, then $h(u,v) = u^2 + u - 2v$, while the expression for $h(u',v')$ remains unchanged, so

$$h(u',v') \leq h(u,v) \iff |u + 2v - 1| \leq 3u - 2v - 1. \tag{2.27}$$

   We split this case into two subcases:

   (a) **Case 2(a)**: If, on the other hand, $u + 2v > 1$, then $|u+2v-1| = u+2v-1$, and (2.27) simplifies to

   $$h(u',v') \leq h(u,v) \iff u + 2v - 1 \leq 3u - 2v - 1 \iff 2v \leq u.$$

   But since $2v - u \leq 0$ by assumption, the result holds in this subcase.

(b) **Case 2(b)**: If $u + 2v \leq 1$, then

$$h(u', v') \leq h(u, v) \iff -u - 2v + 1 \leq 3u - 2v - 1$$

$$\iff u \geq 1/2. \tag{2.28}$$

But of course, the condition $u \geq 1/2$ holds throughout $S^+ \setminus R_1$.

We have shown thus far that $h(u', v') \leq h(u, v)$ if $(u, v) \in S^+ \setminus R_1$ and $M_1(u, v) = (u - 1, u + v - 1)$. It remains to show that $h(u, u + v) \leq h(u, v)$ if $(u, v) \in S^+ \setminus R_1$ and $F_\gamma(u, v) \leq \gamma/2$. By inspection of Figure 1, the intersection of the regions $S^+ \setminus R_1$ and $\{(u, v) : F_\gamma(u, v) \leq \gamma/2\}$ consists of two disjoint sets: $P_1 : \{(u, v) : 0 \leq u + \gamma v \leq \gamma/2, u + 2v \geq 1, u \leq 0\}$, and $P_2 : \{(u, v) : 0 \leq u + \gamma v \leq \gamma/2, u + 2v \leq -1, u \geq 0\}$.

1. **Case 1:** $(u, v) \in P_1$: As $P_1 \subset \{(u, v) : u \leq 0, v \geq 0\}$, the restriction $2v - u \geq 0$ trivially holds, and so $h(u, v) = u^2 + 2v - u$, and

$$h(u, u + v) \leq h(u, v) \iff u^2 + |2(u + v) - u| \leq u^2 + 2v - u$$

$$\iff |2v + u| \leq 2v - u$$

$$\iff 2v + u \leq 2v - u$$

$$\iff u \leq 0$$

which is satisfied by assumption.

2. **Case 2:** $(u, v) \in P_2$: P_2 is contained in the quadrant $\{(u, v) : u \geq 0, v \leq 0\}$, and so $2v - u < 0$, $h(u, v) = u^2 - 2v + u$, and

$$h(u, u+v) \leq h(u, v) \iff |u+2v| \leq u - 2v \iff -(u+2v) \leq u - 2v \iff u \geq 0,$$

which again is satisfied by assumption.

By symmetry of the set R and the map A_γ, we have also that $h(A_\gamma(u, v)) \leq h(u, v)$ if $(u, v) \in S^- \cap \mathbb{R} \setminus R_2$.

Chapter 3

Cross validation in compressed sensing

3.1 Compressed sensing: Redefining traditional notions of sampling

Compressed Sensing (CS) is a fast developing area in applied mathematics, motivated by the reality that most data we store and transmit contains far less information than its dimension suggests. For example, a one-dimensional slice through the pixels in a typical grayscale image will contain segments of smoothly varying intensity, with sharp changes between adjacent pixels appearing only at edges in the image. If a large data vector contains only $k << N$ nonzero entries, or is *k-sparse*, it is common practice to temporarily store the entire vector, possibly with the intent to go back and replace this vector with a smaller dimensional vector encoding the location and

magnitude of its k significant coefficients. In compressed sensing, one instead collects fewer fixed linear measurements of the data to start with, sufficient in number to recover the location and numerical value of the k nonzero coordinates at a later time. Finding "good" linear measurements, as well as fast, accurate, and simple algorithms for recovering the original data from these measurements, are the twofold goals of Compressed Sensing research today.

In the sequel, we restrict our focus to signals that can be represented as real-valued vectors $x = (x_1, x_2, ..., x_N) \in \mathbb{R}^N$. This signal model works well for digital images, which are naturally sparse: for example, a one-dimensional slice through the pixels in a typical grayscale image will contain segments of smoothly varying intensity, with sharp changes between adjacent pixels appearing only at edges in the image. Often this sparsity in information translates into a sparse (or approximately sparse) representation of the data with respect to some standard basis; for the image example, the basis would be a wavelet of curvelet basis. For such N dimensional data vectors that are well approximated by a k-sparse vector (or a vector that contains at most $k << N$ nonzero entries), it is common practice to temporarily store the entire vector, possibly with the intent to go back and replace this vector with a smaller dimensional vector encoding the location and magnitude of its k significant coefficients. In compressed sensing, one instead collects fewer fixed linear measurements of the data to start with, sufficient in number to recover the location and numerical value of the k nonzero coordinates at a later time. Finding 'good' linear measurements, as well as fast, accurate, and simple algorithms for recovering the original data

93

from these measurements, are the twofold goals of compressed sensing research today.

Review of basic CS setup. The data of interest is taken to be a real-valued vector $x \in \mathbb{R}^N$ that is *unknown*, but from which we are allowed up to $m < N$ linear measurements, in the form of inner products of x with m vectors $v_j \in \mathbb{R}^N$ of our choosing. Letting Φ denote the $m \times N$ matrix whose jth row is the vector v_j, this is equivalent to saying that we have the freedom to choose and store an $m \times N$ matrix Φ, along with the m-dimensional measurement vector $y = \Phi x$. Of course, since Φ maps vectors in $\mathbb{R}^N$ to vectors in a smaller dimensional space $\mathbb{R}^m$, the matrix Φ is not invertible, and we thus have no hope of being able to reconstruct an arbitrary N dimensional vector x from such measurements.

However, if the otherwise unknown vector x is specified to be k-sparse, and k is fairly small compared with N, then there do exist matrices Φ for which $y = \Phi x$ uniquely determines x, and allows recovery of x using fast and simple algorithms. It was the interpretation of this phenomenon given by Candes and Tao [18], [17], and Donoho [40], that gave rise to compressed sensing. In particular, these authors define classes of matrices that possess this property. One particularly elegant characterization of this class is via the *Restricted Isometry Property* (RIP) [17]. A matrix Φ with unit normed columns is said to be k-RIP if all singular values of any k column submatrix of Φ lie in the interval $[1 - \delta, 1 + \delta]$ for a given constant $\delta < 1$. For a fixed value of $\delta > 0$, an $m \times N$ matrix Φ whose entries $\Phi_{i,j}$ are independent realizations

of a Gaussian or Bernoulli random variable satisfies $2k$-RIP of level

$$k = K(\delta, m, N) := c_1(\delta)m/\log(N/m)), \text{where } m \leq \frac{1}{2}N, \qquad (3.1)$$

with probability $\geq 1 - 2\exp -c_2(\delta)m$ [6]. Also with high probability, an $m \times N$ matrix obtained by selecting m rows at random from the $N \times N$ discrete Fourier matrix satisfies $2k$-RIP of the same order as (3.1) up to an additional $\log(\log N)$ factor [66]. In fact, the order of k given by (3.1) is optimal given m and N, as shown in [26] using classical results on Gelfand widths of l_1^N unit balls in l_2^N. To date, there exist no deterministic constructions of RIP matrices of this order.

Recovering or approximating x. As shown in [**?**], the following approximation results hold for matrices Φ that satisfy $2k$-RIP with constant $\delta_{2k} \leq 2(3 - \sqrt{2})/7 \approx .4531$:

1. If $x \in \mathbb{R}^N$ is k-sparse, then x can be reconstructed from Φ and the measurement vector $y = \Phi x$ as the solution to the following ℓ_1 minimization problem:

$$x = \mathcal{L}_1(\Phi, y) := \arg\min_{\Phi z = y} \|z\|_1. \qquad (3.2)$$

2. If x is not k-sparse, the error between x and the approximation $\hat{x} = \mathcal{L}_1(\Phi, y)$ is still bounded by

$$\|x - \hat{x}\|_2 \leq \frac{c}{\sqrt{k}} \sigma_k(x)_{l_1^N}, \qquad (3.3)$$

where $c = 2(1 + \delta_{2k})/(1 - \delta_{2k})$, and $\sigma_k(x)_{l_p^N} := \inf_{|z| \leq k} \|x - z\|_{l_p^N}$ denotes the

best possible approximation error in the metric of l_p^N between x and the set of k-sparse signals in $\mathbb{R}^N$. The approximation error $\sigma_k(x)_{l_p^N}$ is realized by the k-sparse vector $x_k \in \mathbb{R}^N$ that corresponds to the vector x with all but the k largest entries set to zero, independent of the l_p^N norm in the approximation $\sigma_k(x)_{l_p^N}$.

This immediately suggests to use the ℓ_1-minimizer $\mathcal{L}_1$ as a means to recover or approximate an unknown x with sparsity constraint. Several other decoding algorithms are used as alternatives to ℓ_1 minimization for recovering a sparse vector x from its image $y = \Phi x$, not because they offer better accuracy (ℓ_1 minimization gives optimal approximation bounds when Φ satisfies RIP), but because they can be faster and easier to implement. For a comprehensive survey on compressed sensing decoding algorithms, we refer the reader to [57].

3.2 Estimating the accuracy of compressed sensing estimates

According to the bound (3.3), the quality of a compressed sensing estimate $\hat{x} = \triangle(\Phi, y)$ depends on how well x can be approximated by a k-sparse vector, where the value of k is determined by the number of rows m composing Φ. While k is assumed to be known and fixed in the compressed sensing literature, no such bound is guaranteed for real-world signal models such as vectors $x \in \mathbb{R}^N$ corresponding to wavelet coefficient sequences of discrete photograph-like images. Thus, the quality of a compressed sensing estimate $\hat{x}$ in general is not guaranteed.

If the error $\|x - \hat{x}\|_2/\|x\|_2$ incurred by a particular approximation $\hat{x}$ were *observed* to be large, then decoding could be repeated using more measurements, perhaps at increasing measurement levels $\{m_1, m_2, ..., m_p\}$, until the error $\|x - \hat{x}_j\|_2/\|x\|_2$ corresponding to m_j measurements were observed to be sufficiently small. Of course, the errors $\|x - \hat{x}_j\|_2$ and $\|x - \hat{x}_j\|_2/\|x\|_2$ are typically not known, as x is unknown. Our main observation in this chapter is that one can apply the Johnson-Lindenstrauss lemma [52] to the set of p points,

$$\{(x - \hat{x}_1), (x - \hat{x}_2), ..., (x - \hat{x}_p)\}. \tag{3.4}$$

In particular, $r = O(\log p)$ measurements of x, provided by $y_\Psi = \Psi x$, when Ψ is, e.g. a Gaussian or Bernoulli random matrix, are sufficient to guarantee that with high probability,

$$4/5\|y_\Psi - \Psi \hat{x}_j\|_2 \leq \|x - \hat{x}_j\|_2 \leq 4/3\|y_\Psi - \Psi \hat{x}_j\|_2 \tag{3.5}$$

and

$$1/3\frac{\|y_\Psi - \Psi \hat{x}_j\|_2}{\|y_\Psi\|_2} \leq \frac{\|x - \hat{x}_j\|_2}{\|x\|_2} \leq 2\frac{\|y_\Psi - \Psi \hat{x}_j\|_2}{\|y_\Psi\|_2} \tag{3.6}$$

for any p compressed sensing estimates. The equivalences (3.5) and (3.6) allow the *measurable* quantities $\|y_\Psi - \Psi \hat{x}_j\|_2$ and $\|y_\Psi - \Psi \hat{x}_j\|_2/\|y_\Psi\|_2$ to function as proxies for the *unknown* quantities $\|x - \hat{x}_j\|_2$ and $\|x - \hat{x}_j\|_2/\|x\|_2$; these proxies can be used to

(a) provide tight numerical upper and lower bounds on the errors $\|x - \hat{x}_j\|_2$ and $\|x - \hat{x}_j\|_2/\|x\|_2$ at up to p compressed sensing estimates $\hat{x}_j$,

97

(b) provide estimates of the underlying k-term approximations $\|x - x_k\|_2$ of x for up to p different values of k, and

(c) return from among a sequence of estimates $(\hat{x}_1, ..., \hat{x}_p)$ with different initialization parameters, an estimate $\hat{x}_{cv} = \arg\min_j \|y_\Psi - \Psi\hat{x}_j\|_2$ having error $\|x - \hat{x}_{cv}\|_2$ that does not exceed a small multiplicative factor of the best possible error in the metric of ℓ_2^N between x and an element from the sequence at hand.

More precisely, all CS decoding algorithms require as input a parameter m corresponding to the number of rows in Φ; some compressed sensing decoding algorithms (such as greedy algorithms) require also a parameter k indicating the sparsity level of x, and other algorithms require as input a bound γ on the expected amount of energy in x outside of its significant coefficients. All CS decoding algorithms can be symbolically represented by functions of the form $\triangle(\Phi, y, k, \gamma)$, and we will give examples where each of the parameters m, k, and γ can be optimized over a sequence of estimates $(\hat{x}_1, \hat{x}_2, ..., \hat{x}_p)$ indexed by increasing hypotheses on each of the parameters m, k, and γ.

The method we propose for parameter selection and error estimation in compressed sensing is reminiscent of *cross validation*, which is a technique used in statistics and learning theory whereby a data set is separated into a training/estimation set and a test/cross validation set, and the test set is used to prevent overfitting on the training set by estimating underlying noise parameters. Indeed, we take a set of m measurements of the unknown x, and use $m - r$ of these measurements, Φx, in a compressed sensing decoding algorithm to return a sequence $(\hat{x}_1, \hat{x}_2, ...)$ of candidate approxima-

tions to x. The remaining r measurements, Ψx, are then used to identify from among this sequence a 'best' approximation $\hat{x} = \hat{x}_j$, along with an estimate of the sparsity level of x. Although the application of cross validation in compressed sensing has been previously proposed by Boufounos, Duarte, and Baraniuk in [11], the context in which it is studied there is different from that of the present paper (we will discuss this difference further in the last section), and in their application one cannot immediately apply the mathematical justification of the Johnson Lindenstrauss lemma that we present below.

3.3 Preliminary notation

Throughout the paper, we will be dealing with large dimensional vectors that have few nonzero coefficients. We use the notation $|x| = n$ to indicate that a vector $x \in \mathbb{R}^N$ has exactly n nonzero coordinates.

We will sometimes use the notation $a \sim_\epsilon b$ as shorthand for the multiplicative relation

$$(1 - \epsilon)a \leq b \leq (1 + \epsilon)a, \tag{3.7}$$

that can be worded as "the quantity a approximates the quantity b to within a multiplicative factor of $(1 \pm \epsilon)$". Note that the relation $\sim_\epsilon$ is not symmetric. Properties of the relation $a \sim_\epsilon b$ are listed below; we omit the proofs, which amount to a string of simple inequalities.

Lemma 3.3.1. *Fix $\epsilon \in (0, 1)$.*

1. *If $a, b \in \mathbb{R}^+$ satisfy $a \sim_\epsilon b$, then $b/[(1+\epsilon)(1-\epsilon)] \sim_\epsilon a$.*

2. *If $a, b, c, d \in \mathbb{R}^+$ satisfy $a \sim_\epsilon b$ and $c \sim_\epsilon d$, then $a/c \sim_\delta b/d$ for parameter $\delta = 2\epsilon/1 - \epsilon$.*

3. *If $(a_1, a_2, ..., a_p)$ and $(b_1, b_2, ..., b_p)$ are sequences in $\mathbb{R}^+$, and $a_j \sim_\epsilon b_j$ for each $1 \leq j \leq p$, then $\min_j a_j \sim_\epsilon \min_j b_j$.*

3.4   Mathematical foundations

The Johnson Lindenstrauss (JL) lemma, in its original form, states that any set of p points in high dimensional Euclidean space can be embedded into $\epsilon^{-2} \log(p)$ dimensions, without distorting the distance between any two points by more than a factor of $(1 \pm \epsilon)$ [52]. In the same paper, it was shown that a random orthogonal projection would provide such an embedding with positive probability. Following several simplifications to the original proof [46], [49], [29], it is now understood that Gaussian random matrices, among other purely random matrix constructions, can substitute for the random projection in the original proof of Johnson and Lindenstrauss. Of the several versions of the lemma now appearing in the literature, the following variant presented in Matousek [55] is most applicable to the current presentation.

Lemma 3.4.1 (Johnson-Lindenstrauss Lemma)**.** *Fix an accuracy parameter $\epsilon \in (0, 1/2]$, a confidence parameter $\delta \in (0, 1)$, and an integer $r \geq r_0 = C\epsilon^{-2} \log \frac{1}{2\delta}$.*

Let $\mathcal{M}$ be a random $r \times N$ matrix whose entries $\mathcal{M}_{i,j}$ are independent realizations of a random variable R that satisfies:

1. $\mathrm{Var}\,[R] = 1/r$ (so that the columns of $\mathcal{M}$ have expected ℓ_2 norm 1)

2. $\mathbb{E}[R] = 0$,

3. For some fixed $a > 0$ and for all λ,

$$Prob[|R| > \lambda] \leq 2e^{-a\lambda^2} \tag{3.8}$$

Then for a predetermined $x \in \mathbb{R}^N$,

$$(1 - \epsilon)\|x\|_{l_2^N} \leq \|\mathcal{M}x\|_{l_2^r} \leq (1 + \epsilon)\|x\|_{l_2^N} \tag{3.9}$$

is satisfied with probability exceeding $1 - \delta$.

The constant C bounding r_0 in Lemma 3.4.1 grows with the parameter a specific to the construction of $\mathcal{M}$ (3.8). Gaussian and Bernoulli random variables R will satisfy the concentration inequality (3.8) for a relatively small parameter a (as can be verified directly), and for these matrices one can take $C = 8$ in Lemma 3.4.1.

The Johnson Lindenstrauss lemma can be made intuitive with a few observations. Since $\mathbb{E}[R] = 0$ and $\mathrm{Var}[R] = \frac{1}{r}$, the random variable $\|\mathcal{M}x\|_2^2$ equals $\|x\|_2^2$ in expected value; that is,

$$\mathbb{E}\big[\,\|\mathcal{M}x\|_2^2\,\big] = \|x\|_2^2. \tag{3.10}$$

101

Additionally, $\|\mathcal{M}x\|_2^2$ inherits from the random variable R a nice concentration inequality:

$$\text{Prob}\left[\|\mathcal{M}x\|_2^2 - \|x\|_2^2 > \epsilon\|x\|_2^2\right] \leq e^{-a(2\epsilon\sqrt{r})^2} \leq \delta/2. \qquad (3.11)$$

The first inequality above is at the heart of the JL lemma; its proof can be found in [55]. The second inequality follows using that $r \geq (2a\epsilon^2)^{-1}\log(\frac{\delta}{2})$ and $\epsilon \leq 1/2$ by construction. A bound similar to (3.11) holds for $\text{Prob}\left[\|\mathcal{M}x\|_2^2 - \|x\|_2^2 < -\epsilon\|x\|_2^2\right]$ as well, and combining these two bounds gives desired result (3.9).

For fixed $x \in \mathbb{R}^N$, a random matrix $\mathcal{M}$ constructed according to Lemma 3.4.1 fails to satisfy the concentration bound (3.9) with probability at most δ. Applying Boole's inequality, $\mathcal{M}$ then fails to satisfy the stated concentration on any of p predetermined points $\{x_j\}_{j=1}^p$, $x_j \in \mathbb{R}^N$, with probability at most $\xi = p\delta$. In fact, a specific value of $\xi \in (0,1)$ may be imposed for fixed p by setting $\delta = \xi/p$. These observations are summarized in the following corollary to Lemma 3.4.1.

Corollary 3.4.2. *Fix an accuracy parameter $\epsilon \in (0, 1/2]$, a confidence parameter $\xi \in (0,1)$, and fix a set of p points $\{x_j\}_{j=1}^p \subset \mathbb{R}^N$. Set $\delta = \xi/p$, and fix an integer $r \geq r_0 = C\epsilon^{-2}\log\frac{1}{2\delta} = C\epsilon^{-2}\log\frac{p}{2\xi}$. If $\mathcal{M}$ is a $r \times N$ matrix constructed according to Lemma 3.4.1, then with probability $\geq 1 - \xi$, the bound*

$$(1-\epsilon)\|x_j\|_{l_2^N} \leq \|\mathcal{M}x_j\|_{l_2^r} \leq (1+\epsilon)\|x_j\|_{l_2^N} \qquad (3.12)$$

obtains for each $j = 1, 2, ..., p$.

3.5   Cross validation in compressed sensing

We return to the situation where we would like to approximate a vector $x \in \mathbb{R}^N$ with an assumed sparsity constraint using $m < N$ linear measurements $y = \mathcal{A}x$ and $m \times N$ matrix $\mathcal{A}$ of our choosing. Continuing the discussion in Section 1, we will not reconstruct x in the standard way by $\hat{x} = \Delta(\mathcal{A}, y, k, \gamma)$ for fixed values of the input parameters, but instead separate the $m \times N$ matrix $\mathcal{A}$ into an $n \times N$ *implementation* matrix Φ and an $r \times N$ *cross validation* matrix Ψ, and separate the measurements y accordingly into y_Φ and y_Ψ. We use the implementation matrix Φ and corresponding measurements y_Φ as input into the decoding algorithm to obtain a sequence of possible estimates $(\hat{x}_1, ..., \hat{x}_p)$ corresponding to increasing one of the input parameters m, k, or γ. We reserve the cross validation matrix Ψ and measurements y_Ψ to estimate each of the error terms $\|x - \hat{x}_j\|_2$ in terms of the computable $\|y_\Psi - \Psi\hat{x}_j\|_2$. We will also estimate the unknown *oracle error*,

$$\eta_{or} = \min_{1 \le j \le p} \|x - \hat{x}_j\|_{l_2^N}, \tag{3.13}$$

corresponding to the best possible approximation to x in the metric of l_2^N from the sequence $(\hat{x}_1, ..., \hat{x}_p)$, using the computable *cross validation error*,

$$\widehat{\eta_{cv}} = \min_{1 \le j \le p} \|\Psi(x - \hat{x}_j)\|_{l_2^r}. \tag{3.14}$$

Our main result, which follows from Corollary (3.4.2), details how the number of cross validation measurements r should be chosen in terms of the desired accuracy ϵ

of estimation, confidence level ξ in the prediction, and number p of estimates $\hat{x}_j$ to be measured.

Theorem 3.5.1. *For a given accuracy $\epsilon \in (0, 1/2]$, confidence $\xi \in (0, 1)$, and number p of estimates $\hat{x}_j \in \mathbb{R}^N$, it suffices to allocate $r = \lceil C\epsilon^{-2} \log \frac{p}{2\xi} \rceil$ rows to a cross validation matrix Ψ of Gaussian or Bernoulli type, normalized according to Lemma 3.4.2 and independent of the estimates $\hat{x}_j$, to obtain with probability greater than or equal to $1 - \xi$, and for each $j = 1, 2, ..., p$, the bounds*

$$\frac{1}{1+\epsilon} \leq \frac{\|x - \hat{x}_j\|_{l_2^N}}{\|\Psi(x - \hat{x}_j)\|_{l_2^\ell}} \leq \frac{1}{1-\epsilon} \tag{3.15}$$

and

$$\frac{1 - 3\epsilon}{(1+\epsilon)(1-\epsilon)^2} \leq \frac{\|x - \hat{x}_j\|_{l_2^N}/\|x\|_{l_2^N}}{\|\Psi(x - \hat{x}_j)\|_{l_2^\ell}/\|\Psi x\|_{l_2^\ell}} \leq \frac{1}{(1-\epsilon)^2}$$

$$\tag{3.16}$$

and also

$$\frac{1}{1+\epsilon} \leq \frac{\eta_{or}}{\widehat{\eta_{cv}}} \leq \frac{1}{1-\epsilon}. \tag{3.17}$$

Proof. .

- The bounds in (3.15) are obtained by application of Lemma 3.4.2 to the p points $u_j = x - \hat{x}_j$, and rearranging the resulting bounds according to Lemma

3.3.1 part (1). The bound (3.17) follows from the bounds (3.15) and part (3) of Lemma 3.3.1.

- The bounds in (3.16) are obtained by application of Lemma 3.4.2 to the $p+1$ points $u_0 = x, u_j = x - \hat{x}_j$, and regrouping the resulting bounds according to part (2) of Lemma 3.3.1.

$\square$

Remark 3.5.2. The measurements making up the cross validation matrix Ψ must be independent of the measurements comprising the rows of the implementation matrix Φ. This comes from the requirement in Lemma 3.4.1 that the matrix Ψ be independent of the points $u_j = x - \hat{x}_j$. This requirement is crucial, as observed when $\hat{x}$ solves the ℓ_1 minimization problem

$$\hat{x} = \arg \min_{z \in \mathbb{R}^N} \|z\|_1 \text{ subject to } \Phi z = \Phi x, \tag{3.18}$$

in which case the constraint $\Phi(\hat{x} - x) = 0$ clearly precludes the rows of Φ from giving any information about the error $\|\hat{x} - x\|_2$.

Remark 3.5.3. If the full compressed sensing matrix $\mathcal{A} = [\Phi \, ; \, \Psi]$ is itself of Gaussian or Bernoulli type, then one can obtain a more accurate approximation to the unknown quantities $\|x - \hat{x}_j\| / \|x\|$ by using all of the measurements $\mathcal{A}x$ in the estimates $\|\Psi(x - \hat{x}_j)\|_{l_2^\ell} / \|\mathcal{A}x\|_{l_2^m}$.

Remark 3.5.4. Proposition 3.5.1 should be applied with a different level of care depending on what information about the sequence $(x - x_1, x - x_2, ..., x - x_p)$ is sought.

If the minimizer $\hat{x} = \arg\min_{1 \le j \le p} \|\Psi(x - \hat{x}_j)\|_{l_2^r}$ is sufficient for one's purposes, then the precise normalization of Ψ in Proposition 3.5.1 is not important. The normalization doesn't matter either for estimating the normalized quantities $\|x - \hat{x}_j\|_2 / \|x\|_2$. On the other hand, if one is using cross validation to obtain estimates for the quantities $\|x - \hat{x}_j\|_2$, then normalization is absolutely crucial, and one must observe the normalization factor given by Lemma 3.4.2 that depends on the number of rows r allocated to the cross validation matrix Ψ.

3.6 Applications

3.6.1 Estimation of the best k-term approximation error

We have already seen that if the $m \times N$ matrix Φ satisfies $2k$-RIP with parameter $\delta \le\approx .45$, and $\hat{x} = \mathcal{L}_1(\Phi, \Phi x)$ is returned as the solution to the ℓ_1 minimization problem (4.3.1), then the error between x and the approximation $\hat{x}$ is bounded by

$$\|x - \hat{x}\|_2 \le \frac{c}{\sqrt{k}} \sigma_k(x)_{l_1^N}. \tag{3.19}$$

Several other decoding algorithms in addition to ℓ_1 minimization enjoy the reconstruction guarantee (3.19) under similar bounds on Φ, such as the Iteratively Reweighted Least Squares algorithm (IRLS) [32], and the greedy algorithms CoSAMP [57] and Subspace Pursuit [27]. It has recently been shown [78] [38] that if the bound (3.19) is obtained, and if $x - \hat{x}$ lies in the null space of Φ (as is the case for the decoding algorithms just mentioned), then if Φ is a Gaussian or a Bernoulli random matrix,

the error $\|x - \hat{x}\|_2$ also satisfies a bound, with high probability on Φ, with respect to the ℓ_2^N residual, namely

$$\|x - \hat{x}\|_2 \leq c\sigma_k(x)_{l_2^N}, \tag{3.20}$$

for a reasonable constant c depending on the RIP constant δ_{2k} of Φ. In the event that (3.20) is obtained, a cross validation estimate $\|\Psi(x - \hat{x})\|_{l_2^r}$ can be used to lower bound the residual $\sigma_k(x)_{l_2^N}$, with high probability, according to

$$(1 - \epsilon)\|\Psi(x - \hat{x})\|_{l_2^\ell} \leq \|x - \hat{x}\|_{l_2^N} \leq c\sigma_k(x)_{l_2^N}, \tag{3.21}$$

with $O(\frac{1}{\epsilon^2})$ rows reserved for the matrix Ψ (3.5.1). At this point, we will use Corollary 3.2 of [58], where it is proved that if the bound (3.19) holds for $\hat{x}$ with constant c, then the same bound will hold for

$$\hat{x}_k = \arg \min_{z:|z| \leq k} \|\hat{x} - z\|_{l_2^N}, \tag{3.22}$$

the best k-sparse approximation to $\hat{x}$, with constant $\tilde{c} = 3c$. Thus, we may assume without loss of generality that $\hat{x}$ is k-sparse, in which case $\|\Psi(x - \hat{x})\|_{l_2^r}$ also provides an upper bound on the residual $\sigma_k(x)_{l_2^N}$ by

$$(1 + \epsilon)\|\Psi(x - \hat{x})\|_{l_2^r} \geq \|x - \hat{x}\|_{l_2^N} \geq \sigma_k(x)_{l_2^N}. \tag{3.23}$$

With almost no effort then, cross validation can be incorporated into many decoding algorithms to obtain tight upper and lower bounds on the unknown k-sparse approximation error $\sigma_k(x)_{l_2^N}$ of x. More generally, the allocation of $10 \log p$ measurements

107

to the cross validation matrix Ψ is sufficient to estimate the errors $(\|x - x_{k_j}\|_2)_{j=1}^p$

or the normalized approximation errors $(\|x - x_{k_j}\|_2 / \|x\|_2)_{j=1}^p$ at p sparsity levels k_j

by decoding p times, adding m_j measurements to the implementation matrix Φ_j at

each repetition. Recall that the quantities k_j and m_j are related according to (3.1).

3.6.2 Choice of the number of measurements m

Photograph-like images have wavelet or curvelet coefficient sequences $x \in \mathbb{R}^N$ that

are *compressible* [37] [16], having entries that obey a power law decay

$$|x|_{(k)} \leq c_s k^{-s}, \tag{3.24}$$

where $x_{(k)}$ denotes the kth largest coefficient of x in absolute value, the parameter

$s > 1$ indicates the level of compressibility of the underlying image, and c_s is a

constant that depends only on s and the normalization of x. From the definition

(3.24), compressible signals are immediately seen to satisfy

$$\|x - x_k\|_1 / \sqrt{k} \leq c_s' k^{-s+1/2}, \tag{3.25}$$

so that the solution $\hat{x}_m = \mathcal{L}_1(\Phi, \Phi x)$ to the ℓ_1 minimization problem (4.3.1) using an

$m \times N$ matrix Φ of optimal RIP order $k = cm/\log(N/m))$ satisfies

$$\|x - \hat{x}_m\|_2 \leq c_{s,\delta} k^{-s+1/2}. \tag{3.26}$$

The number of measurements m needed to obtain an estimate $\hat{x}_m$ satisfying $\|x - \hat{x}_m\|_2 \leq \tau$ for a predetermined threshold τ will vary according to the compressibility of the image at hand. Armed with a total of m measurements, the following decoding method that *adaptively* chooses the number of measurements for a given signal x presents a more democratic alternative to standard compressed sensing decoding structure:

Table 3.1: *CS decoding structure with adaptive number of measurements*

1. *Input*: The m-dimensional vector $y = \Phi x$, the $m \times N$ matrix Φ, (in some algorithms) the sparsity level k, and (again, in some algorithms) a bound γ on the noise level of x, the number p of row subsets of Φ, $(\Phi_1, \Phi_2, ..., \Phi_p)$, corresponding to increasing number of rows $m_1 < m_2 < ... < m_p < m$, and threshold $\tau > 0$.

2. *Initialize* the decoding algorithm at $j = 1$.

3. *Estimate* $\hat{x}_j = \triangle(\Phi_{m_j}, y_{m_j}, k, \gamma)$ with the decoder $\triangle$ at hand, using only the first m_j measurement rows of Φ. The previous estimate $\hat{x}_{j-1}$ can be used for "warm initialization" of the algorithm, if applicable. The remaining $r_j = m - m_j$ measurement rows are allocated to a cross validation matrix Ψ_j that is used to estimate the resulting error $\|x - \hat{x}_j\|_2 / \|x\|_2$.

4. *Increment* j by 1, and iterate from step 3 if stopping rule is not satisfied.

5. *Stop*: at index $j = j^* < p$ if $\|x - x_{m_j}\|_2 / \|x\|_2 \leq \tau$ holds with near certainty, as indicated by

$$\frac{\sqrt{r_j} \|\Psi(x - x_{m_j})\|_2 / \|\Phi x\|_2}{\sqrt{r_j} - 3 \log p} \leq \tau \qquad (3.27)$$

according to Proposition 3.5.1. If the maximal number of decoding measurements $m_p < m$ have been used at iteration p, and (3.27) indicates that $\|x - \hat{x}_{m_p}\|_2 / \|x\|_2 > \tau$ still, return $\hat{x}_m = \triangle(\Phi, y, k, \gamma)$ using all m measurements, but with a warning that the underlying image x is probably too dense, and its reconstruction is not trustworthy.

109

3.6.3 Choice of regularization parameter in homotopy-type algorithms

Certain compressed sensing decoding algorithms iterate through a sequence of intermediate estimates $\hat{x}_j$ that could be potential optimal solutions to x under certain reconstruction parameter choices. This is the case for greedy and homotopy-continuation based algorithms. In this section, we study the application of cross validation to the intermediate estimates of decoding algorithms of homotopy-continuation type.

LASSO is the name coined in [71] for the problem of minimizing of the following convex program:

$$\hat{x}^{[\tau]} = \arg \min_{\|z\|_1 \leq \tau} \|\Phi x - \Phi z\|_{\ell_2} \tag{3.28}$$

The two terms in the LASSO optimization problem (3.28) enforce data fidelity and sparsity, respectively, as balanced by the regularization parameter τ. In general, choosing an appropriate value for τ in (3.28) is a hard problem; when Φ is an underdetermined matrix, as is the case in compressed sensing, the function $f(\tau) = \|x - \hat{x}^{[\tau]}\|_2$ is unknown to the user but is seen empirically to have a minimum at a value of τ in the interval $[0, \|\Phi x\|_\infty]$ that depends on the unknown noise level and/or and compressibility level of x.

The *homotopy continuation* algorithm [53], which can be viewed as the appropriate variant of LARS [53], is one of many algorithms for solving the LASSO problem

(3.28) at a predetermined value of τ; it proceeds by first initializing τ' to a value sufficiently large to ensure that the ℓ_1 penalization term in (3.28) completely dominates the minimization problem and $x^{[\tau']} = 0$ trivially. The homotopy continuation algorithm goes on to generate $x^{[\tau']}$ for decreasing τ' until the desired level for τ is reached. If $\tau = 0$, then the homotopy method traces through the entire solution path $x^{[\tau']} \in \mathbb{R}^N$ for $\tau' \geq 0$ before reaching the final algorithm output $x^{[0]} = \mathcal{L}_1(\Phi, y)$ corresponding to the ℓ_1 minimizer (4.3.1).

From the non-smooth optimality conditions for the convex functional (3.28), it can be shown that the solution path $\hat{x}^{[\tau]} \in \mathbb{R}^N$ is a piecewise-affine function of τ [53], with "kinks" possible only at a finite number of points $\tau \in \{\tau_1, \tau_2, ...\}$. Proposition 3.5.1 suggests a method whereby an appropriate value of $\tau*$ can be chosen from among a subsequence of the kinks $(\tau_1, \tau_2, ..., \tau_p)$ by solving the minimization problem $\hat{x}^{[\tau*]} = \arg\min_{j \leq p} \|\Psi(x - \hat{x}^{[\tau_j]})\|_2$ for appropriate cross validation matrix Ψ. Moreover, since the solution $x - \hat{x}^{[\tau]}$ for $\tau_j \leq \tau \leq \tau_{j+1}$ is restricted to lie in the two-dimensional subspace spanned by $x - \hat{x}^{[\tau_j]}$ and $x - \hat{x}^{[\tau_{j+1}]}$, one can combine the Johnson Lindenstrauss Lemma with a covering argument analogous to that used to derive the RIP property for Gaussian and Bernoulli random matrices in [6], to cross validate the entire *continuum* of solutions $\hat{x}^{[\tau]}$ between $\tau_1 \leq \tau \leq \tau_p$. More precisely, the following bound holds under the conditions outlined in Theorem (3.5.1), with the exception that $2r$ (as opposed to r) measurements are reserved to Ψ:

$$\frac{1}{1+\epsilon} \leq \frac{\min_{\tau_1 \leq \tau \leq \tau_p} \|x - \hat{x}^{[\tau]}\|_2}{\min_{\tau_1 \leq \tau \leq \tau_p} \|\Psi(x - \hat{x}^{[\tau]})\|_2} \leq \frac{1}{1-\epsilon} \tag{3.29}$$

Unfortunately, the bound (3.29) is not strong enough to *provably* evaluate the entire solution path $\hat{x}^{[\tau]}$ for $\tau \geq 0$, because the best upper bound on the number of kinks on a generic LASSO solution path can be very large. One can prove that this number is bounded by 3^N, by observing that if $\hat{x}^{[\tau_1]}$ and $\hat{x}^{[\tau_2]}$ have the same sign pattern, then $\hat{x}^{[\tau]}$ also has the same sign pattern for $\tau_1 \leq \tau \leq \tau_2$. Applying Proposition 3.5.1 to $p = 3^N$ points $x - \hat{x}_j$, this suggests that $O(N)$ rows would need to be allocated to a cross validation matrix Ψ in order for Proposition 3.5.1 and the corollary (3.29) to apply to the entire solution path, which clearly defeats the compressed sensing purpose. However, whenever the matrix Φ is an $m \times N$ compressed sensing matrix of random Gaussian, Bernoulli, or partial Fourier construction, it is observed *empirically* that the number of kinks along a homotopy solution path behaves like $O(m)$ for generic vectors $x \in \mathbb{R}^N$ used to generate the path, see Figure 3.1. This suggests, at least heuristically, that the allocation of $O(\log m)$ out of m compressed sensing measurements of this type suffices to ensure that the error $\|x - \hat{x}^{[\tau]}\|_2$ for the solution $\hat{x}^{[\tau]} = \arg\min_{\tau \geq 0} \|\Psi(x - \hat{x}^{[\tau]})\|_2$ will be within a small multiplicative factor of the best possible error in the metric of ℓ_2^N obtainable by any approximant $\hat{x}^{[\tau]}$ along the solution curve $\tau \geq 0$. At the value of τ corresponding to $\hat{x}^{[\tau]}$, the LASSO solution (3.28) can be computed using all m measurements Φ as a final approximation to x.

The *Dantzig selector* (DS) [19] refers to a minimization problem that is similar
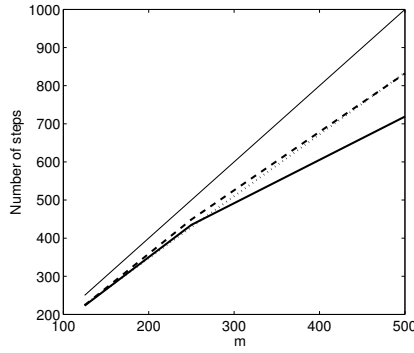
Figure 3.1: Number of steps in the homotopy continuation algorithm, as described in Section 3.6.3, as a function of the number of rows m comprising the encoding matrix Φ. The thick solid, dashed, and dotted lines represent a bound satisfied by 95 out of 100 trials corresponding to independent realizations of an $m \times N$ Gaussian random matrix Φ and N-dimensional Gaussian random vector x, when $N = 500, 1000$, and 2000, respectively. The lines are generated from data at $m = 125, 250$, and 500. The number of steps appears to depend sublinearly on the number of rows m and does not seem to depend on the ambient dimension N. The thin solid line corresponds to the observed upper bound of $2m$.

in form to the LASSO problem:

$$\hat{x}_\tau = \arg \min_{z \in \mathbb{R}^N} \|\Phi x - \Phi z\|_{\ell_\infty} + \tau \|z\|_1 \qquad (3.30)$$

The difference between the DS (3.30) and LASSO (3.28) is the choice of norm (ℓ_∞ versus ℓ_2) on the fidelity-promoting term. Homotopy-continuation based algorithms have also been developed to solve the minimization problem (3.30) by tracing through the solution path $\hat{x}_{\tau'}$ for $\tau' \geq \tau$. As the minimization problem (3.30) can be reformulated as a linear program, its solution path $\hat{x}_\tau \in \mathbb{R}^N$ is seen to be a piecewise *constant* function of τ, in contrast to the LASSO solution path. In practice, the total number of breakpoints $(\tau_1, \tau_2, ...)$ in the domain $0 \leq \tau$ is observed to be on the same order of magnitude as m when the $m \times N$ matrix Φ satisfies RIP [50]; thus, the procedure just described to cross validate the LASSO solution path can be adapted to cross validate the solution path of (3.30) as well.

Thus far we have not discussed the possibility of using cross validation as a stopping criterion for homotopy-type decoding algorithms. Along the LARS homotopy curve (3.28), most of the breakpoints $(\tau_1, \tau_2, ...)$ appear only near the end of the curve in a very small neighborhood of $\tau = 0$. These breakpoints incur only miniscule changes in the error $\|x - \hat{x}_{\tau_j}\|_2$ even though they account for most of the computational expense of the LARS decoding algorithm. Therefore, it would be interesting to adapt such algorithms, perhaps using cross validation, to stop once τ^* is reached for which the error $\|x - \hat{x}_{\tau^*}\|_2$ is sensed to be sufficiently small.

3.6.4 Choice of sparsity parameter in greedy-type algorithms

Greedy compressed sensing decoding algorithms also iterate through a sequence of intermediate estimates $\hat{x}_j$ that could be potential optimal solutions to x under certain reconstruction parameter choices. *Orthogonal Matching Pursuit* (OMP), which can be viewed as the prototypical greedy algorithm in compressed sensing, picks columns from the implementation matrix Φ one at a time in a greedy fashion until, after k iterations, the k-sparse vector $\hat{x}_k$, a linear combination of the k columns of Φ chosen in the successive iteration steps, is returned as an approximation to x. The OMP algorithm is listed in Table 2. Although we will not describe the algorithm in full detail, a comprehensive study of OMP can be found in [72].

Table 3.2: *Orthogonal Matching Pursuit Basic Structure*

1. *Input*: The m-dimensional vector $y = \mathcal{B}x$, the $m \times N$ encoding matrix Φ whose j^{th} column is labeled ϕ_j, and the sparsity bound k.

2. *Initialize* the decoding algorithm at $j = 1$, the residual $r_0 = y$, and the index set $\Lambda_0 = \emptyset$.

3. *Estimate*

   (a) Find an index λ_j that realizes the bound $(\Phi^T r_{j-1})_{\lambda_j} = \|\Phi^T r_{j-1}\|_\infty$.

   (b) Update the index set $\Lambda_j = \Lambda_{j-1} \cup \lambda_j$ and the submatrix of contributing columns: $\Phi_j = [\Phi_{j-1},\ \phi_{\lambda_j}]$

   (c) Update the residual:

   $$
   \begin{aligned}
   s_j &= \ \arg\min_x \|\Phi_j x - y\|_2 = (\Phi_j^T \Phi_j)^{-1} \Phi_j^T y, \\
   a_j &= \ \Phi_j x_j \\
   r_j &= \ r_{j-1} - a_j.
   \end{aligned}
   $$

   (d) The estimate $\hat{x}_j$ for the signal has nonzero indices at the components listed in Λ_j, and the value of the estimate $\hat{x}_j$ in component λ_i equals the *ith* component of s_j.

4. *Increment* j by 1 and iterate from step 3, if $j < k$.

5. *Stop*: at $j = k$. Output $\hat{x}_{omp} = \hat{x}_k$ as approximation to x.

Note in particular that OMP requires as input a parameter k corresponding to the expected sparsity level for $x \in \mathbb{R}^N$. Such input is typical among greedy algorithms in compressive sensing (in particular, we refer the reader to [57], [32], and [27]). As shown in [72], OMP will recover with high probability a vector x having at most $k \leq m/\log N$ nonzero coordinates from its image Φx if Φ is a (known) $m \times N$ Gaussian or Bernoulli matrix with high probability. Over the more general class of vectors $x = x_d + \mathcal{N}$ that can be decomposed into a d-sparse vector x_d (with d

presumably less than or equal to k) and additive noise vector $\mathcal{N}$, we might expect an intermediate estimate $\hat{x}_s$ to be a better estimate to x than the final OMP output $\hat{x}_k$, at least when $d << k$. Assuming that the signal x admits a decomposition of the form $x = x_d + \mathcal{N}$, the sequence of intermediate estimates $(\hat{x}_1, ..., \hat{x}_k)$ of an OMP algorithm can be cross validated in order to estimate the noise level and recover a better approximation to x. We will study this particular application of cross validation in more detail below.

3.7 Orthogonal matching pursuit: a case study

As detailed in Table 2, a single index λ_j is added to a set Λ_j estimated as the j most significant coefficients of x at each iteration j of OMP; following the selection of Λ_j, an estimate $\hat{x}_j$ to x is determined by the least squares solution,

$$\hat{x}_j = \arg \min_{\text{supp}(z) \in \Lambda_j} \|\Phi z - y\|_2, \tag{3.31}$$

among the subspace of vectors $z \in \mathbb{R}^N$ having nonzero coordinates in the index set Λ_j. OMP continues as such, adding a single index λ_j to the set Λ_j at iteration j, until $j = k$ at which point the algorithm terminates and returns the k-sparse vector $\hat{x}_{omp} = \hat{x}_k$ as approximation to x.

Suppose x has only d significant coordinates. If d could be specified beforehand, then the estimate $\hat{x}_d$ at iteration $j = d$ of OMP would be returned as an approximation to x. However, the sparsity d is not known in advance, and k will instead be

an upper bound on d. As the estimate $\hat{x}_j$ in OMP can be then identified with the hypothesis that x has j significant coordinates, the application of cross-validation as described in the previous section applies in a very natural way to OMP. In particular, we expect $\hat{x}_{or}$ and $\hat{x}_{cv}$ of Proposition 3.5.1 to be close to the estimate $\hat{x}_j$ at index $j = |x|$ corresponding to the true sparsity of x; furthermore, in the case that $|x|$ is significantly less than k, we expect the cross validation estimate $\hat{x}_{cv}$ to be a better approximation to x than the OMP-returned estimate $\hat{x}_k$. We will put this intuition to the test in the following numerical experiment.

3.7.1 Experimental setup

We initialize a signal x_0 of length $N = 3600$ and sparsity level $d = 100$ as

$$x_0(j) = \begin{cases} 1, & \text{for } j = 1...100 \\ 0, & \text{else.} \end{cases} \tag{3.32}$$

Noise is then added to $x_a = x_0 + \mathcal{N}_a$ in the form of a Gaussian random variable $\mathcal{N}_a$ distributed according to

$$\mathcal{N}_a \sim N(0, .05), \tag{3.33}$$

and the resulting vector x_a is renormalized to satisfy $\|x_a\|_{l_2^N} = 1$. This yields an expected noise level of

$$\mathbb{E}\big[\sigma_d(x_a)\big] \approx .284. \tag{3.34}$$

We fix the input $k = 200$ in Table 2, and assume we have a total number of com-

pressed sensing measurements $m = 800$. A number r of these m measurements are allotted to cross validation, while the remaining $n = m - r$ measurements are allocated as input to the OMP algorithm in Table 2. This experiment aims to numerically verify Proposition 3.5.1; to this end, we specify a confidence $\xi = 1/100$, and solve for the accuracy ϵ according to the relation $r = \epsilon^{-2} \log(\frac{k}{2\xi})$; that is,

$$\epsilon(r) = \sqrt{\frac{\log(\frac{k}{2\xi})}{r}} \approx \frac{3}{\sqrt{r}}. \tag{3.35}$$

Note that the specification (3.35) corresponds to setting the constant $C = 1$ in Proposition 3.5.1. Although $C \geq 8$ is needed for the proof of the Johnson Lindenstrauss lemma at present, we find that in practice $C = 1$ already upper bounds the optimal constant needed for Proposition 3.5.1 for Gaussian and Bernoulli random ensembles.

A single (properly normalized) Gaussian $n \times N$ measurement matrix Φ is generated (recall that $n = m$ - r) , and this matrix and the measurements $y = \Phi x$ are provided as input to the OMP algorithm; the resulting sequence of estimates $(\hat{x}_1, \hat{x}_2, ..., \hat{x}_k)$ is stored. The final estimate $\hat{x}_k$ from this sequence is the returned OMP estimate $\hat{x}_{omp}$ to x. The error $\eta_{omp} = \|\hat{x}_{omp} - x\|_2$ is greater than or equal to the oracle error of the sequence, $\eta_{or} = \min_{\hat{x}_j} \|x - \hat{x}_j\|_2$.

With the sequence $(\hat{x}_1, \hat{x}_2, ..., \hat{x}_k)$ at hand, we consider 1000 realizations Ψ_q of an $r \times N$ cross validation matrix having the same componentwise distribution as Φ, but normalized to have variance $1/r$ according to Theorem 3.4.1. The cross validation

error

$$\widehat{\eta_{cv}}(q) = \min_j \|\Psi_q(x - \hat{x}_j)\|_{l_2^r} \tag{3.36}$$

is measured at each realization Ψ_q; we plot the average $\overline{\widehat{\eta_{cv}}}$ of these 1000 values and intervals centered at $\overline{\widehat{\eta_{cv}}}$ having length equal to twice the empirical standard deviation. Note that we are effectively testing 1000 trials of OMP-CV, the algorithm which modifies OMP to incorporate cross validation so that $(\hat{x}_{cv}, \widehat{\eta_{cv}})$ are output instead of $\hat{x}_{omp} = \hat{x}_k$.

At the specified value of ξ, Proposition 3.5.1 part (3.16) (with constant $C = 1$) implies that

$$(1 - \epsilon)\eta_{or} \le \widehat{\eta_{cv}}(q) \le (1 + \epsilon)\eta_{or} \tag{3.37}$$

should obtain on at least 990 of the 1000 estimates $\widehat{\eta_{cv}}(q)$; in other words, at least 990 of the 1000 discrepancies $|\eta_{or} - \widehat{\eta_{cv}}(q)|$ should be bounded by

$$0 \le |\widehat{\eta_{cv}}(q) - \eta_{or}| \le \epsilon\eta_{or}. \tag{3.38}$$

Using the relation (3.35) between ϵ and r, this bound becomes tighter as the number r of CV measurements increases; however, at the same time, the oracle error η_{or} increases with r for fixed m as fewer measurements $n = m - r$ are input to OMP. An ideal number r of CV measurements should not be too large or too small; Figure 1 suggests that setting aside just enough measurements r such that $\epsilon \le .6$ is satisfied in (3.35) serves as a good heuristic to choose the number of cross validation measurements (in Figure 1, $\epsilon \le .6$ is satisfied by taking only $r = 30$ measurements).

We indicate the theoretical bound (3.37) with dark gray in Figure 1, which is compared to the interval in light gray of the 990 values of $\eta_{cv}(q)$ that are closest to η_{or} in actuality.

This experiment is run for several values of r within the interval $[5, 90]$; the results are plotted in Figure 1(a), with the particular range $r \in [5, 30]$ blown up in Figure 1(b).

We have also carried out this experiment with a smaller noise variance; i.e. $x_b = x_0 + \mathcal{N}_b$ is subject to additive noise

$$\mathcal{N}_b \sim N(0, .02). \tag{3.39}$$

The signal x_b is again renormalized to satisfy $\|x_b\|_{l_2^N} = 1$; it now has an expected noise level of

$$\mathbb{E}\big[\sigma_d(x_b)\big] \approx .116. \tag{3.40}$$

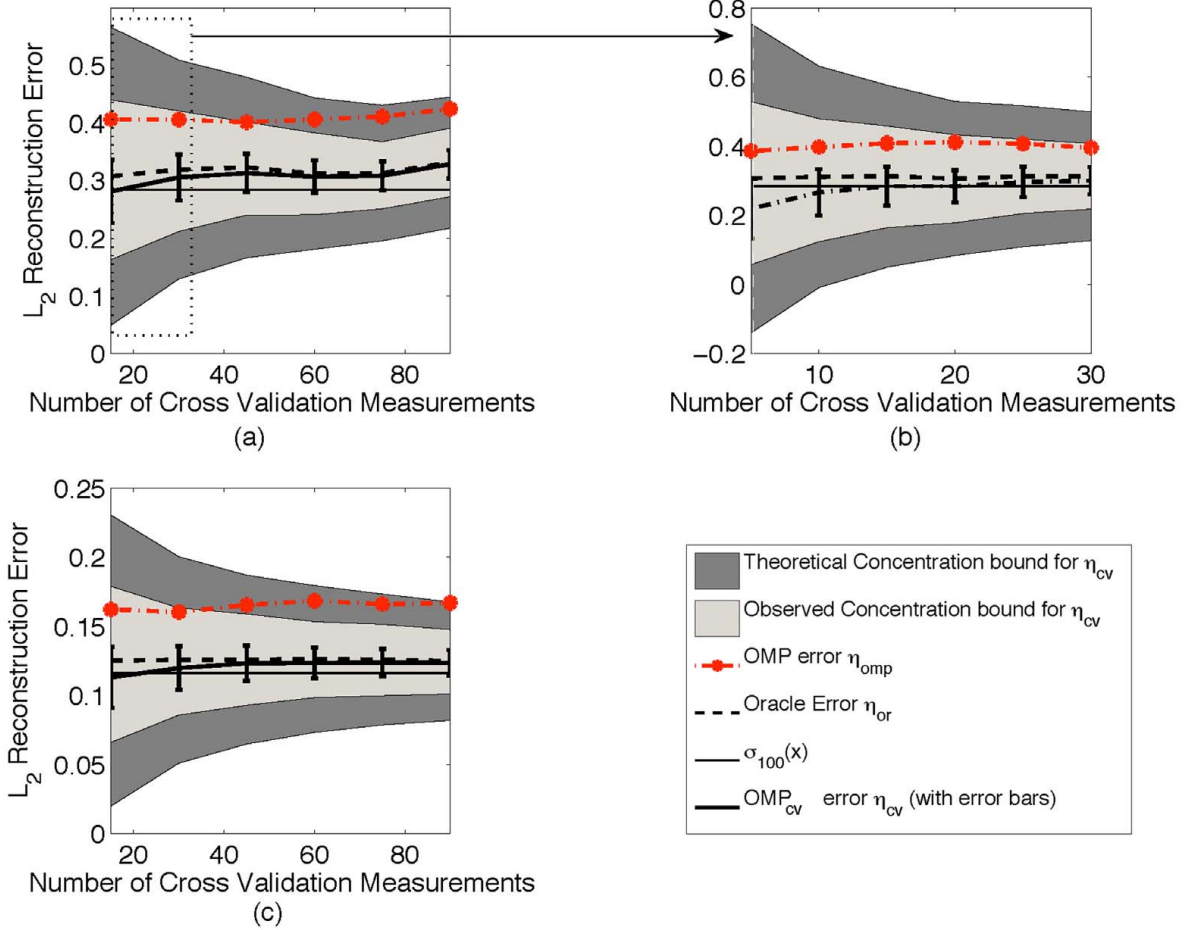The results of this experiment are plotted in Figure 1(c).

Figure 3.2: Comparison of the reconstruction algorithms OMP and OMP-CV. We fix the parameters $N = 3600, m = 800, k = 200$, and underlying sparsity $d = 100$, but vary the number r of the total m measurements reserved for cross validation from 5 to 90, using the remaining $n = m - r$ measurements for training. The underlying signal has residual $\sigma_{100}(x) \approx .284$ in Figures 1(a) and 1(b), and $\sigma_{100}(x) \approx .116$ in Figure 1(c), as shown for reference by the thin horizontal line. In both cases, the OMP-CV error η_{cv} (the solid black line with error bars; each point represents the average of 1000 trials) gives a better approximation to the residual error than does OMP (dot-dashed line) with very high probability, even when as few as 20 of the total 800 measurements are used for cross validation. Even though η_{cv} is guaranteed to provide a tighter bound for η_{or} as the number r of CV measurements increases, at the same time, the oracle error η_{or} becomes a worse indicator of the residual $\sigma_{100}(x)$ because fewer measurements $n = m - r$ are input to OMP.

3.7.2 Experimental results

1. We remind the reader that the cross-validation estimates $\widehat{\eta_{cv}}$ are observable to the user, while the values of η_{omp}, η_{or}, along with the noise level $\sigma_d(x)$, are not available to the user. Nevertheless, $\widehat{\eta_{cv}}$ can serve as a proxy for η_{or} according to (3.37), and this is verified by the plots in Figure 1. $\widehat{\eta_{cv}}$ can also provide an upper bound on $\sigma_d(x)$, as is detailed in Section 5.1.

2. The theoretical bound (3.37) is seen to be tight, when compared with the observed concentration bounds in Figure 1.

3. With high probability, the estimate $\hat{x}_{cv}(15)$ using $r = 15$ out of the alloted $m = 800$ measurements will be a better estimate of x than the OMP estimate: $\|\hat{x}_{cv}(15) - x\|_2 \leq \|\hat{x}_{omp}(15) - x\|_2$. With *overwhelming probability*, the estimate $\hat{x}_{cv}(30)$ will result in error $\|\hat{x}_{cv}(30) - x\|_2 \leq \|\hat{x}_{omp}(30) - x\|_2$. We note that the estimates $\hat{x}_{cv}(15)$ and $\hat{x}_{cv}(30)$ correspond to accuracy parameters $\epsilon(15) = .8405$ and $\epsilon(30) = .5943$ in (3.35), indicating that $\epsilon \leq .6$ is a good heuristic to determine when enough CV measurements have been reserved.

4. The OMP-CV estimate $\hat{x}_{cv}$ will have more pronounced improvement over the OMP estimate $\hat{x}_{omp}$ when there is larger discrepancy between the true sparsity d of x_0 and the upper bound k used by OMP (in Figure (1), $d = 100$ and $k = 200$). In contrast, OMP-CV will not outperform OMP in approximation accuracy when d is close to k; however, the multiplicative relation (3.37) guarantees that OMP-CV will not underperform OMP, either.

3.8 Beyond compressed sensing

The Compressed Sensing setup can be viewed within the more general class of *under-determined linear inverse problems*, in which $x \in \mathbb{R}^N$ is to be reconstructed from a known $m \times N$ underdetermined matrix $\mathcal{A}$ and lower dimensional vector $y = \mathcal{A}x$ using a decoding algorithm $\Delta : \mathbb{R}^m \to \mathbb{R}^N$; in this broader context, $\mathcal{A}$ is given to the user, but not necessarily *specified by* the user as in compressed sensing. In many cases, a prior assumption of sparsity is imposed on x, and an iterative decoding algorithm for solving the LASSO problem (3.28) will be used to reconstruct x from y [30]. If it is possible to take on the order of $r = \log p$ additional measurements of x by an $r \times N$ matrix Ψ satisfying the conditions of Lemma 3.4.1, then all of the analysis presented in this paper applies to this more general setting. In particular, the error $\|x - \hat{x}_j\|_{l_2^N}$ at up to $j \leq p$ successive approximations $\hat{x}_j$ of the decoding algorithm Δ may be bounded from below and above using the quantities $\|\Psi(x - \hat{x}_j)\|_{\ell_2^r}$, and the final approximation $\hat{x}$ to x can be chosen from among the entire sequence of esti-mates $\hat{x}_j$ as outlined in Proposition 3.5.1; an earlier estimate $\hat{x}_j$ may approximate x better than a final estimate $\hat{x}_p$ which contains the artifacts of parameter overfitting occurring at later stages of iteration.

3.9 Extensions

We have presented an alternative approach to compressed sensing in which a certain number r of the m allowed measurements of a signal $x \in \mathbb{R}^N$ are reserved to track the error in decoding by the remaining $m-r$ measurements, allowing us to choose a best

approximation to x in the metric of ℓ_2^N out of a sequence of p estimates $(\hat{x}_j)_{j=1}^p$, and estimate the error between x and its best approximation by a k-sparse vector, again with respect to the metric of ℓ_2^N. We detailed how the number r of such measurements should be chosen in terms of desired accuracy ϵ of estimation, confidence level ξ in the prediction, and number p of decoding iterations to be measured; in general, $r = O(\log(p))$ measurements suffice. Several important issues remain unresolved; we mention only a few below.

1. **Noisy measurements.** Consider the noisy measurement model,

$$y = \mathcal{A}x + \mathcal{N}, \tag{3.41}$$

where $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$, and $\mathcal{N} \sim N(0, \sigma^2)$ is a Gaussian random variable that accounts for both noise and quantization error on the measurements $\mathcal{A}x$. Because measurement noise and quantization error are unavoidable in any real-world sensing device, any proposed compressed sensing technique should extend to the model (3.41). Cross validation is studied in [11] in the context of noisy measurements, assuming that x is truly sparse and that $\mathcal{N} \sim N(0, \sigma^2)$ is Gaussian noise. The experimental results in [11] indicate that cross validation works well in the presence of noise, and Proposition 3.5.1 provides intuition as to why. Suppose that $x \in \mathbb{R}^N$ is k-sparse, so that the best possible reconstruction to x using n noisy measurements $y_1 = \Phi x + \mathcal{N}_1$ of the m total noisy measurements

$y = \mathcal{A}x + \mathcal{N}$ is bounded by the ℓ_2 norm of $\mathcal{N}_1$:

$$\|x - \hat{x}\|_{\ell_2^N} \sim \mathbb{E}\big[\|\mathcal{N}_1\|_{\ell_2^n}\big] = \sigma\sqrt{n}. \tag{3.42}$$

Alternatively, the noise vector $\mathcal{N}_{cv}$ corresponding to the remaining $r = m - n$ cross validation measurements has *exponentially* smaller expected ℓ_2 norm $\mathbb{E}[\|\mathcal{N}_{cv}\|_{\ell_2^r}] = \sigma\sqrt{10\log n}$, assuming that $r = 10\log n$. It follows by application of the triangle inequality,

$$
\begin{aligned}
(1-\epsilon)\|x - \hat{x}\|_{\ell_2^r} - \sigma\sqrt{\log n} &\leq \|\Psi(x - \hat{x}) + \mathcal{N}_{cv}\|_{\ell_2^r} \\
&\leq (1+\epsilon)\|x - \hat{x}\|_{\ell_2^r} + \sigma\sqrt{\log n},
\end{aligned}
$$

whence the approximation error $\|x - \hat{x}\|_{\ell_2^r} \sim \sigma\sqrt{n}$ dominates both lower and upper bounds on the noisy cross validation error.

2. **Other cross validation techniques.** The cross validation technique promoted in this paper corresponds in particular to the technique of *holdout cross validation* in statistics, where a data set is partitioned into a single training and cross validation set (as a rule of thumb, the cross validation set is usually taken to be less than or equal to a third of the size of the training set; in the the current paper, we have shown that the Johnson Lindenstrauss lemma provides a theoretical justification of how many, or, more precisely, how few, cross validation measurements are needed in the context of compressed sensing). Other forms of cross validation, such as *repeated random subsampling* cross validation

or *K-fold* cross validation, remain to be analyzed in the context of compressed sensing. The former technique corresponds to repeated application of holdout cross validation, with r cross validation measurements out of the total m measurements chosen by random selection at each application. The results are then averaged (or otherwise combined) to produce a single estimation. The latter technique, K-fold cross validation, also corresponds to repeated application of holdout cross validation. In this case, the m measurements are partitioned into K subsets of equal size r, and cross-validation is repeated exactly K times with each of the K subsets of measurements used exactly once as the validation set. The K results are again combined to produce a single estimation. Although Proposition 3.5.1 does not directly apply to these cross validation models, the experimental results of Section 6 suggest that, equiped with an $m \times N$ matrix satisfying the requirements of Lemma 3.4.1, the application of K fold cross validation to subsets of the measurements of size $r << m - r$ just large enough that $\epsilon > 0$ in Proposition 3.5.1 for fixed accuracy ξ and constant $C = 1$ can be combined to accurately approximate the underlying signal with near certainty.

3. **Cross validation in different reconstruction metrics.** We have only considered cross validation over the metric of ℓ_2. However, the error $\|x - \hat{x}\|_{\ell_2^N}$, or *root mean squared error*, is just one of several metrics used in image processing for analyzing the quality of a reconstruction $\hat{x} \in \mathbb{R}^N$ to a (known) image $x \in \mathbb{R}^N$. In fact, the ℓ_1 reconstruction error $\|x - \hat{x}\|_{\ell_1^N}$ has been argued to outperform the root mean squared error as an indicator of reconstruction quality [37]. Unfortunately, Proposition 3.5.1 cannot be extended to the metric

of ℓ_1, as there exists no ℓ_1 analog of the Johnson Lindenstrauss Lemma [25]. However, it remains to understand the extent to which cross validation in compressed sensing can be applied over a broader class of image reconstruction metrics, perhaps using more refined techniques than those considered in this paper.

4. **Cross validation with more general CS matrix ensembles.** Many more compressed sensing matrices in addition to those satisfying the requirements of Lemma 3.4.1 can be used for cross validation purposes, in the sense that a concentration bound similar to (3.4.1) will hold for *most* input x. Indeed, it has been recently shown [39] that such random partial Fourier matrices, if multiplied on the right by an $N \times N$ diagonal matrix of independent and identically distributed Bernoulli random variables, will satisfy the concentration bound (3.4.1) with slightly larger number of measurements $r \geq \epsilon^{-2} \log^3 \frac{1}{2\delta}$.

Chapter 4

Free discontinuity problems meet iterative thresholding

4.1 Introduction

Free-discontinuity problems describe situations where the solution of interest is defined by a function and a lower dimensional set consisting of the discontinuities of the function. Hence, the derivative of the solution is assumed to be a 'small' function almost everywhere except on sets where it concentrates as a singular measure. This is the case, for instance, in crack detection from fracture mechanics or in certain digital image segmentation problems. If we discretize such situations for numerical purposes, the free-discontinuity problem in the discrete setting can be re-formulated as that of finding a derivative vector with small components at all but a few entries that exceed a certain threshold. This problem is similar to those encountered

in Compressed sensing, where vectors with a small number of dominating components in absolute value are recovered from a few given linear measurements via the minimization of related energy functionals. Several iterative thresholding algorithms that intertwine gradient-type iterations with thresholding steps have been designed to recover sparse solutions in this setting. It is natural to wonder if and/or how such algorithms can be used towards solving discrete free-discontinuity problems. The current chapter explores this connection, and, by establishing an iterative thresholding algorithm for discrete free-discontinuity problems, provides new insights on properties of minimizing solutions thereof.

4.1.1   Free-discontinuity problems: the Mumford-Shah functional

The terminology 'free-discontinuity problem' was introduced by De Giorgi [36] to indicate a class of variational problems that consist in the minimization of a functional, involving both volume and surface energies, depending on a closed set $K \subset \mathbb{R}^d$, and a function u on $\mathbb{R}^d$ usually smooth outside of K. In particular,

- K is not fixed a priori and is an unknown of the problem;

- K is not a boundary in general, but a free-surface inside the domain of the problem.

The best-known example of a free-discontinuity problem is the one modelled by the so-called Mumford-Shah functional [56], which is defined by

$$J(u, K) := \int_{\Omega \setminus K} \left[|\nabla u|^2 + \alpha(u - g)^2\right] dx + \beta \mathcal{H}^{d-1}(K \cap \Omega).$$

The set Ω is a bounded open subset of $\mathbb{R}^d$, $\alpha, \beta > 0$ are fixed constants, and $g \in L^\infty(\Omega)$. Here $\mathcal{H}^N$ denotes the N-dimensional Hausdorff measure. Throughout this paper, the dimension of the underlying Euclidean space $\mathbb{R}^d$ will always be $d = 1$ or $d = 2$. In the context of visual analysis, g is a given noisy image that we want to approximate by the minimizing function $u \in W^{1,2}(\Omega \setminus K)$; the set K is simultaneously used in order to *segment* the image into connected components. For a broad overview on free-discontinuity problems, their analysis, and applications, we refer the reader to [3].

If the set K were fixed, then the minimization of J with respect to u would be a relatively simple problem, equivalent to solving the following system of equations:

$$\begin{aligned} \Delta u &= \alpha(u - g), &&\text{in } \Omega \setminus K, \\ \frac{\partial u}{\partial \nu} &= 0, &&\text{on } \partial\Omega \cup K, \end{aligned}$$

where ν is the outward-pointing normal vector at any $x \in \partial\Omega \cup K$. Therefore the relevant unknown in free-discontinuity problems is the set K. Ensuring the existence of minimizers (u, K) of J is a challenging problem because there is no topology on the closed sets that ensures

(a) compactness of minimizing sequences and

(b) lower semicontinuity of the Hausdorff measure.

Indeed, it is well-known, by the direct method of calculus of variations [28, Chapter 1], that the two previous conditions ensure the existence of minimizers. However, the problem becomes more manageable if we restrict our domain to functions $u \in BV(\Omega) \cap W^{1,2}(\Omega \setminus K)$, and make the identification $K \equiv \overline{S_u}$ where S_u is the well-defined discontinuity set of u. In this case, we need to work only with a topology on the space $BV(\Omega)$ of bounded variation, and no set topology is anymore required.

Unfortunately the space $BV(\Omega)$ is 'too large'; it contains Cantor-like functions whose approximate gradient vanishes, $\nabla u = 0$, almost everywhere, and whose discontinuity set has measure zero, $\mathcal{H}^{d-1}(S_u) = 0$. As these functions are dense in $L^2(\Omega)$, the problem is trivialized; see [3] for details.

Nevertheless, it is possible to give a meaningful formulation of the functional J if we exclude such functions and restrict J to the space $SBV(\Omega)$ constituted of BV-functions with vanishing Cantor part. If we assume again $K \equiv \overline{S_u}$, the solution can be recast as the minimization of

$$\mathcal{J}(u) = \int_{\Omega \setminus S_u} \left[|\nabla u|^2 + \alpha(u - g)^2 \right] dx + \beta \mathcal{H}^{d-1}(S_u). \qquad (4.1)$$

The existence of minimizers in SBV for the functional (4.1) was established by Ambrosio on the basis of his fundamental compactness theorem in [2], see also [3,

Theorem 4.7 and Theorem 4.8].

4.1.2   Γ-convergence approximation to free-discontinuity problems

The discontinuity set S_u of a SBV-function u is not an object that can be easily handled, especially numerically. This difficulty gave rise to the development of approximation methods for the Mumford-Shah functional and its minimizers where *sets* are no longer involved, and instead substituted by suitable *indicator functions*. In order to understand the theoretical basis for these approximations, we need to introduce the notion of Γ-convergence, which is today considered one of the most successful notions of 'variational convergence'; we state only the definition of Γ-convergence below, but refer the reader to [28, 13] for a broad introduction.

Definition 4.1.1. *Let (X, d) be a metric space[1] and let $f, f_n : X \rightarrow [0, \infty]$ be functions for $n \in \mathbb{N}$. We say that $(f_n)_{n \in \mathbb{N}}$ Γ-converges to f if the following two conditions are satisfied:*

i) for any sequence $(x_n)_n \subset X$ converging to x,

$$\liminf_n f_n(x_n) \geq f(x);$$

---

[1]Observe that by [28, Proposition 8.7] suitable bounded sets X endowed with the weak topology induced by a larger Banach space are indeed metrizable, so this condition is not that restrictive.

ii) for any $x \in X$, there exists a sequence $(x_n)_n \subset X$ converging to x such that

$$\limsup_n f_n(x_n) \leq f(x).$$

One important consequence of Definition 4.1.1 is that if a sequence of functionals f_n Γ-converges to a target functional f, then the corresponding minimizers of f_n also converge to minimizers of f, see [28, Corollary 7.30].

We define now

$$F_\varepsilon(u, v) := \int_\Omega \left[v^2 |\nabla u|^2 + \alpha (u - g)^2 \right] + \frac{\beta}{2} \left(\varepsilon |\nabla v|^2 + \frac{(1 - v)^2}{\varepsilon} \right) dx \qquad (4.2)$$

over the domain $(L^2(\Omega))^2$, along with the related functional

$$\mathcal{J}_\varepsilon(u, v) := \begin{cases} F_\varepsilon(u, v) & \text{, if } v \in W^{1,2}(\Omega),\ uv \in W^{1,2}(\Omega),\ \text{and } 0 \leq v \leq 1, \\ \infty & \text{, else.} \end{cases} \qquad (4.3)$$

Note that at the minimizer (u, v) of $\mathcal{J}_\varepsilon$, the function $0 \leq v \leq 1$ tends to indicate the discontinuity set S_u of the functional (4.1) as $\varepsilon \to 0$. In [4] Ambrosio and Tortorelli proved the following Γ-approximation result:

Theorem 4.1.2 (Ambrosio-Tortorelli '90)**.** *For any infinitesimal sequence $(\varepsilon_n)_n$, the*

functional $\mathcal{J}_{\varepsilon_n}(u, v)$ Γ-converges in $(L^2(\Omega))^2$ to the functional

$$\mathcal{J}(u, v) := \begin{cases} \mathcal{J}(u) & , \text{ if } v \equiv 1, \\ \infty & , \text{ otherwise.} \end{cases} \tag{4.4}$$

4.1.3   Discrete approximation

In fact, the Mumford-Shah functional is the continuous version of a previous discrete formulation of the image segmentation problem proposed by Geman and Geman in [47]; see also the work of Blake and Zisserman in [9]. Let us recall this discrete approach. For simplicity let $d = 2$ (as for image processing problems), $\Omega = [0, 1]^2$, and let $u_{i,j} = u(hi, hj)$, $(i, j) \in \mathbb{Z}^2$ be a discrete function defined on $\Omega_h := \Omega \cap h\mathbb{Z}^2$, for $h > 0$. Define $W_h(t) = \min\{t^2, \beta/h\}$ to be the truncated quadratic potential, and

$$\begin{aligned} \mathcal{J}_{\sqrt{\beta/h}}(u) \ &:= \ h^2 \sum_{(hi,hj)\in\Omega_h} W_h\left(\frac{u_{i+1,j} - u_{i,j}}{h}\right) \\ &+ \ h^2 \sum_{(hi,hj)\in\Omega_h} W_h\left(\frac{u_{i,j+1} - u_{i,j}}{h}\right) \\ &+ \ \alpha h^2 \sum_{(hi,hj)\in\Omega_h} (u_{i,j} - g_{i,j})^2. \end{aligned} \tag{4.5}$$

Chambolle [20, 21] gave formal clarification as to how the discrete functional $\mathcal{J}_{\sqrt{\beta/h}}$ approximates the continuous functional $\mathcal{J}$ of Ambrosio: discrete sequences can be interpolated by piecewise linear functions in such a way as to allow for discontinuities when the discrete finite differences of the sampling values are large enough. On the basis of this identification of discrete functions on Ω_h and functions defined on the 'continuous domain' Ω, we have the following result:

Theorem 4.1.3 (Chambolle '95). *The functional $\mathcal{J}_{\sqrt{\beta/h}}$ Γ-converges in $\mathcal{B}(\Omega)$ (the space of Borel-measurable functions, which is metrizable, see [21] for details) to*

$$\mathcal{J}^{cab}(u) = \int_{\Omega\setminus S_u} \left[|\nabla u|^2 + \alpha(u-g)^2\right] dx + \beta\mathcal{C}(S_u),$$

as $h \to 0$, where $\mathcal{C}$ is the so-called 'cab-driver' measure defined below.

Basically $\mathcal{C}$ measures the length of a curve only through its projections along horizontal and vertical axes; for a regular C^1 curve $c = \gamma([0,1])$, with $\gamma(t) = (\gamma_1(t), \gamma_2(t)) \in \Omega$, we have

$$\mathcal{C}(c) = \int_0^1 \left(|\gamma_1'(t)| + |\gamma_2'(t)|\right) dt.$$

The reason this anisotropic (or, direction dependent) measure appears, in place of the Hausdorff measure in the Mumford-Shah functional, is due to the approximation of derivatives by finite differences defined on a 'rigid' squared geometry. A discretization of derivatives based on meshes *adapted* to the morphology of the discontinuity indeed leads to precise approximations of the Mumford-Shah functional [22, 12].

4.1.4   Free-discontinuity problems and discrete derivatives

In the literature, several methods have been proposed to numerically approximate minimizers of the Mumford-Shah functional [8, 12, 20, 21, 54]. In particular, a relaxation algorithm, based essentially on alternated minimization of a finite element approximation of the Ambrosio and Tortorelli functional (4.3), leads to iterated so-

lutions of suitable elliptic PDEs, where the differential part includes the auxiliary variable v which encodes and indicates information about the discontinuity set. These implementations are basically finite dimensional approximations to the following algorithm: *Starting with $v^{(0)} \equiv 1$, iterate*

$$
\begin{cases}
u^{(n+1)} := \arg\min_{u \in W^{1,2}(\Omega)} \mathcal{J}_\varepsilon(u, v^{(n)}) \\
v^{(n+1)} := \arg\min_{v \in W^{1,2}(\Omega)} \mathcal{J}_\varepsilon(u^{(n+1)}, v).
\end{cases}
$$

However, neither has a proof of convergence of this iterative process to its stationary points been explicitly provided in the literature, nor have the properties of such stationary points been investigated, especially in case of genuine inverse problems (see the discussion in Subsection 4.1.4).

In this paper, we take a different approach and investigate how minimization of the Γ-approximating discrete functionals (4.5) can be implemented efficiently by iterative thresholding on the discrete derivatives. Unlike the aforementioned approach, we will be able to provide a rigorous proof of convergence to stationary points, which coincide with local minimizers of the discrete Mumford-Shah functional. Moreover, we are able to characterize stability properties of such stationary points, and demonstrate the stability of global minimizers of the discrete Mumford Shah functional.

Let us recall: the solutions u of a free-discontinuity problem are supposed to be smooth out of a minimal ipersurface K. This means that the distributional derivative of u is a 'small function' everywhere except on K where it coincides with a singular measure. In the discrete approximation (4.5), the vector of finite differences $(w_j) = (\frac{u_{i,j+1}-u_{i,j}}{h}, \frac{u_{i+1,j}-u_{i,j}}{h})$ corresponds to a piecewise constant function that is

137

small everywhere except for a few locations, corresponding to $|w_j| \geq \sqrt{\beta/h}$, that approximate the discontinuity set K. So, in terms of derivatives, solutions of (4.5) are vectors having only few large entries. In the next section, we clarify how we can indeed work with just derivatives and forget the primal problem.

The 1-D case

Let us assume for simplicity that the dimension $d = 1$, the domain $\Omega = [0, 1]$, and the parameters $\alpha = \beta = 1$. Denote by $u_i = u(hi)$ a discrete function defined on $hi \in \Omega_h := \Omega \cap h\mathbb{Z}$, for $h > 0$; note that the vector $(u_i) \in \mathbb{R}^n$ for $n = \lfloor 1/h \rfloor$. In this setting, the discrete functional (4.5) reduces to

$$\mathcal{J}_{\sqrt{1/h}}(u) = h \sum_{(hi) \in \Omega_h} W_h \left(\frac{u_{i+1} - u_i}{h} \right)$$
$$+ h \sum_{(hi) \in \Omega_h} (u_i - g_i)^2,$$

where we recall that $W_h(t) = \min\{t^2, 1/h\}$. Since no geometrical anisotropy is now involved ($d = 1$), it is possible to show that this discrete functional Γ-converges precisely to the corresponding Mumford-Shah functional on intervals [20].

For $(u_i)_{hi \in \Omega_h}$ we define the discrete derivative as the matrix $D_h : \mathbb{R}^n \to \mathbb{R}^{n-1}$ that

138

maps $(u_i)_{hi\in\Omega_h}$ into $\left(\frac{u_{i+1}-u_i}{h}\right)_i$, given by

$$
D_h = \frac{1}{h}
\begin{pmatrix}
-1 & 1 & 0 & \ldots & \ldots & 0 \\
0 & -1 & 1 & 0 & \ldots & 0 \\
\ldots & & & & & \\
0 & 0 & \ldots & \ldots & -1 & 1
\end{pmatrix}.
\tag{4.6}
$$

It is not too difficult to show that

$$
u = D_h^\dagger D_h u + c,
$$

where $D_h^\dagger$ is the pseudo-inverse matrix of D_h (in the Moore-Penrose sense; note that $D_h^\dagger$ maps $\mathbb{R}^{n-1}$ into $\mathbb{R}^n$ and is an injective operator) and c is a constant vector which depends on u, and the values of its entries coincide with the mean value $h\sum_{hi\in\Omega_h} u_i$ of u. Therefore, any vector u is uniquely identified by the pair $(D_h u, c)$.

Since constant vectors comprise the null space of D_h, the orthogonality relation $\langle D_h^\dagger D_h u, c\rangle_{\ell_2^n} = 0$ holds for any vector u and any constant vector c. Here the scalar product $\langle\cdot,\cdot\rangle_{\ell_2^n} = \sum_i u_i v_i$ is the standard Euclidean scalar product, which induces the Euclidean norm $\|u\|_{\ell_2^n} := \left(\sum_i u_i^2\right)^{1/2}$. Using this orthogonality property, we have that

$$
\begin{aligned}
\|u-g\|_{\ell_2^n}^2 &= \|D_h^\dagger D_h u - D_h^\dagger D_h g + (c - c_g)\|_{\ell_2^n}^2 \\
&= \|D_h^\dagger D_h u - D_h^\dagger D_h g\|_{\ell_2^n}^2 + \|c - c_g\|_{\ell_2^n}^2
\end{aligned}
$$

139

Hence, with a slight abuse of notation, we can reformulate the original problem in terms of derivatives, and mean values, by

$$\mathcal{J}_{1/\sqrt{h}}(z,c) \;=\; h\|D_h^\dagger z - f\|_{\ell_2^n}^2 + h\|c - c_g\|_{\ell_2^n}^2 + h\sum_i \min\left\{|z_i|^2, \frac{1}{h}\right\}$$

where $z = D_h u$ and $f = D_h^\dagger D_h g$. Of course at the minimizer u we have $c = c_g$, since this term in $\mathcal{J}_{1/\sqrt{h}}$ does not depend on z. Therefore, $\|c - c_g\|_2^2$ does not play any role in the minimization and can be neglected. Once the minimal derivative vector z is computed, we can assemble the minimal u by incorporating the mean value of g as follows:

$$u = D_h^\dagger z + c_g.$$

The 2-D case, discrete Schwartz conditions, and constrained optimization

Let us assume now $d = 2$, $\Omega = [0,1]^2$, and again $\alpha = \beta = 1$. Denote $u_{i,j} = u(hi, hj)$, $(i,j) \in \mathbb{Z}^2$, a discrete function defined on $\Omega_h := \Omega \cap h\mathbb{Z}^2$, $n = \lfloor 1/h \rfloor$, and

$$
\begin{aligned}
\mathcal{J}_{1/\sqrt{h}}(u) \;\; := \;\; & h^2 \sum_{(hi,hj)\in\Omega_h} W_h\left(\frac{u_{i+1,j} - u_{i,j}}{h}\right) \\
+ \;\; & h^2 \sum_{(hi,hj)\in\Omega_h} W_h\left(\frac{u_{i,j+1} - u_{i,j}}{h}\right) \\
+ \;\; & h^2 \sum_{(hi,hj)\in\Omega_h} \left(u_{i,j} - g_{i,j}\right)^2.
\end{aligned}
$$

In two dimensions, we have to consider the derivative matrix $D_h : \mathbb{R}^{n^2} \to \mathbb{R}^{2n(n-1)}$ that maps the vector $(u_{j+(i-1)n}) := (u_{i,j})$ to the vector composed of the finite differences in the horizontal and vertical directions u_x and u_y respectively, given by

$$D_h u := \begin{bmatrix} u_x \\ u_y \end{bmatrix}, \quad \begin{cases} (u_x)_{j+n(i-1)} := (u_x)_{i,j} := \frac{u_{i+1,j}-u_{i,j}}{h}, i = 1, \ldots, n-1, j = 1, \ldots, n \\ (u_y)_{j+(n-1)(i-1)} := (u_y)_{i,j} := \frac{u_{i,j+1}-u_{i,j}}{h}, i = 1, \ldots, n, j = 1, \ldots, n-1 \end{cases}.$$

Note that its range $R(D_h) \subset \mathbb{R}^{2n(n-1)}$ is a $(n^2 - 1)$-dimensional subspace because $D_h c = 0$ for constant vectors $c \in \mathbb{R}^{n^2}$. Again, we have the differentiation-integration formula, given by

$$u = D_h^\dagger D_h u + c,$$

where $D_h^\dagger$ is the pseudo-inverse matrix of D_h (in the Moore-Penrose sense); note that $D_h^\dagger$ maps $R(D_h)$ injectively into $\mathbb{R}^{n^2}$. Also, c is a constant vector that depends on u, and the values of its entries coincide with the mean value $h^2 \sum_{(hi,hj) \in \Omega_h} u_{i,j}$ of u.

Proceeding as before and again with a slight abuse of notation, we can reformulate the original discrete functional (4.5) in terms of derivatives, and mean values, by

$$\mathcal{J}_{1/\sqrt{h}}(z,c) = h^2 \left[\|D_h^\dagger z - f\|_{\ell_2^{n^2}}^2 + \|c - c_g\|_{\ell_2^{n^2}}^2 + \sum_{i,j} \min \left\{ |z_{i,j}|^2, \frac{1}{h} \right\} \right].$$

where $z = D_h u \in \mathbb{R}^{2n(n-1)}$, and $f = D_h^\dagger D_h g \in \mathbb{R}^{n^2}$. Of course $c = c_g$ is again assumed at the minimizer u, since this latter term in $\mathcal{J}_{1/\sqrt{h}}$ does not depend on z.

141

However, in order to minimize only over vectors in $\mathbb{R}^{2n(n-1)}$ that are derivatives of vectors in $\mathbb{R}^{n^2}$, we must minimize $\mathcal{J}_{1/\sqrt{h}}(z,c)$ subject to the constraint $D_h D_h^\dagger z = z$.
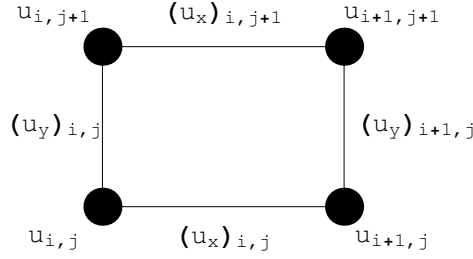

Figure 4.1: Compatibility conditions of derivatives in 2D.

The $2n(n-1)$ linearly independent constraints $D_h D_h^\dagger z = z$ are equivalent to the *discrete Schwartz constraints*[2],

$$(u_y)_{i,j} + (u_x)_{i,j+1} = (u_y)_{i+1,j} + (u_x)_{i,j}, \tag{4.7}$$

that establish the equivalence of the length of the paths from $u_{i,j}$ to $u_{i+1,j+1}$, whether one moves in vertical first and then in horizontal direction or in horizontal first and then in vertical direction (see Figure 4.1).

In short, we arrive at the following constrained optimization problem:

---

[2]These discrete conditions correspond to the well-known Schwartz mixed derivative theorem for which $\partial_{xy} u = \partial_{yx} u$ for any $u \in C^2(\Omega)$.

$$
\begin{cases}
\text{Minimize} \quad \mathcal{J}_{1/\sqrt{h}}(z) = h^2 \big[\|Tz - f\|_{\ell_2^{n^2}}^2 + \sum_{i,j} \min\{|z_{i,j}|^2, \tfrac{1}{h}\} \big]. \\
\\
\text{subject to} \quad \mathcal{Q}z = 0,
\end{cases}
\tag{4.8}
$$

for $T = D_h^\dagger$ and $\mathcal{Q} = \mathcal{I} - D_h D_h^\dagger$. Once the minimal derivative vector z is computed, we can assemble the minimal u by incorporating the mean value of g as follows:

$$
u = D_h^\dagger z + c_g.
$$

Regularization of inverse problems by means of the Mumford-Shah constraint

The Mumford-Shah regularization term

$$
MS(u) = \int_{\Omega \setminus S_u} |\nabla u|^2 + \beta \mathcal{H}^{d-1}(S_u),
\tag{4.9}
$$

has been used frequently in inverse problems for image processing [41, 62], such as inpainting and tomographic inversion. Despite the successful numerical results observed in the aforementioned papers for the minimization of functionals of the type

$$
\mathcal{J}(u) = \alpha \|Ku - g\|_{L^2(\Omega)} + MS(u),
\tag{4.10}
$$

where $K : L^2(\Omega) \to L^2(\Omega)$ is a bounded operator which is not boundedly invertible, no rigorous results on existence of minimizers are currently available in the literature.

Indeed, the Ambrosio compactness theorem [2] used for the proof of the case $K = I$ does not apply in general. A few attempts towards using the regularization MS for inverse problems in fracture detection appear in the work of Rondi [63, 64, 65], although restrictive technical assumptions on the admissible discontinuities of the solutions are required.

As one of the contributions to this paper, we show that discretizations of regularized functionals of the type (4.10) *always* have minimizers (see Theorem 4.2.2). More precisely, these discretizations correspond to functionals of the form,

$$
\mathcal{J}_{\sqrt{\beta/h}}(u) := \alpha h^2 \|Ku - g\|_{\ell_2}^2 + h^2 \sum_{(hi,hj) \in \Omega_h} \left[W_h \left(\frac{u_{i+1,j} - u_{i,j}}{h} \right) + W_h \left(\frac{u_{i,j+1} - u_{i,j}}{h} \right) \right].
$$
(4.11)

and we prove that such functionals admit minimizers. Note that the discrete Mumford-Shah approximation (4.5) can be written in this form. We go on to show that such minimizers can be characterized by certain fixed point conditions, see Theorem 4.6.1 and Theorem 4.6.2. As a consequence of these achievements we can prove that global minimizers are always isolated, although not necessarily unique, whereas local minimizers may constitute a continuum of unstable equilibria. Hence, our analysis will shed light on fundamental properties, virtues, and limitations, of regularization by means of the Mumford-Shah functional MS, and provide a rigorous justification of the numerical results appearing in the literature.

It is useful to show how the discrete functional (4.11) can be still expressed in terms

of the sole derivatives for general K. As done before in the case $K = \mathcal{I}$, and with the now usual identification $u = (D_h u, c)$, we can rewrite the functional in terms of derivatives and mean value as follows:

$$\mathcal{J}_{\sqrt{\beta/h}}(z, c) = h^2\alpha\|KD_h^\dagger z - (g - Kc)\|_2^2 + h^2\sum_{i,j}\min\left\{|z_{i,j}|^2, \frac{\beta}{h}\right\}, \quad (4.12)$$

Note that in general we cannot anymore split orthogonally the discrepancy $\|KD_h^\dagger z - (g - Kc)\|_2^2$ into a sum of two terms which depend only on derivatives z and mean value c respectively. Nevertheless, for fixed z, it is straightforward to show that $\bar{c} = \arg\min_c \mathcal{J}_{\sqrt{\beta/h}}(z, c)$ depends on z via an affine map. Indeed we can compute

$$\bar{c} = \left(\frac{\langle K\mathbf{1}, g - KD_h^\dagger z\rangle}{\|K\mathbf{1}\|_{\ell^2}^2}\right)\mathbf{1},$$

where $\mathbf{1}$ is the constant vector with entries identically 1. Here we assume that $\mathbf{1} \notin \ker K$, that is a necessary condition in order to be able to identify the mean value of minimizers (a similar condition is required anytime we deal with regularization functionals which depend on the sole derivatives, see, e.g., [23, 73]). By substituting this expression for $\bar{c}$ into (4.12), it is clear that the minimization of functionals (4.11) can be reformulated, in terms of the sole derivatives, as constrained minimization problems of the form (4.8).

4.2 Existence of minimizers for a class of discrete free-discontinuity problems

In light of the observations above, we can transform the problem of the minimization of functionals of the type (4.10), by means of discretization first and then reduction to sole derivatives, into the (possibly, but not necessarily) constrained minimization problem:

$$
\begin{cases}
\text{Minimize} \quad \mathcal{J}_r(u) = \left[\|Tu - g\|_{\ell_2^M}^2 + \sum_{i=1}^N \min\{|u_i|^2, r^2\} \right]. \\
\text{subject to} \quad \mathcal{Q}u = 0.
\end{cases}
\tag{4.13}
$$

Our first result ensures the existence of minimizers for the constrained optimization problem (4.13):

Proposition 4.2.1. *Assume $r > 0$, and fix linear operators $T : \mathbb{R}^N \to \mathbb{R}^M$ and $\mathcal{Q} : \mathbb{R}^N \to \mathbb{R}^{M'}$, which are identified in the following with their matrices with respect to the canonical bases. We also fix $g \in \mathbb{R}^M$. The constrained minimization problem*

$$
\begin{cases}
\text{Minimize} \quad \mathcal{J}_r(u) = \left[\|Tu - g\|_{\ell_2^M}^2 + \sum_{i=1}^N \min\{|u_i|^2, r^2\} \right] \\
\text{subject to} \quad \mathcal{Q}u = 0.
\end{cases}
\tag{4.14}
$$

has minimizers u^*.

Proof. We begin by noting that $\inf_{\mathcal{Q}u=0} \mathcal{J}_r(u)$ is well-defined and finite, since $\mathcal{J}_r \geq 0$ is bounded from below. It remains to show that there exists a vector u^* that satisfies $\mathcal{J}_r(u^*) = \inf_{u \in \mathbb{R}^N} \mathcal{J}_r(u)$. Towards this goal, consider the following partition

$\mathcal{P} = \{\mathcal{U}_{\mathcal{I}_j}\}_{j=1}^{2^N}$ of $\mathbb{R}^N$ indexed by the subsets $\mathcal{I}_j$ of the index set $\mathcal{I} = \{1, 2, ..., N\}$, as follows:

$$\mathcal{U}_{\mathcal{I}_j} := \{u \in \mathbb{R}^N : |u_i| \le r, i \in \mathcal{I}_j, |u_i| > r, i \in \mathcal{I}/\mathcal{I}_j\}. \tag{4.15}$$

The minimization of $\mathcal{J}_r$ subject to $\mathcal{Q}u = 0$ and constrained to the closure of the subset $\mathcal{U}_{\mathcal{I}_j}$ can be reformulated as a quadratic optimization problem, for which the classical Frank-Wolfe theorem [7] guarantees the existence of a minimizer $u(\mathcal{I}_j)$. Now, since $\mathbb{R}^N = \cup_j \mathcal{I}_j$, the minimal value of $\mathcal{J}_r$ subject to $\mathcal{Q}u = 0$ and over all of $\mathbb{R}^N$ is just the minimal value from the finite set $\{\mathcal{J}_r(u(\mathcal{I}_j)) : \quad j = 1, \dots, 2^N\}$; that is,

$$\min_{\mathcal{Q}u=0} \mathcal{J}_r(u) \quad = \quad \min_{\mathcal{I}_j \subset \mathcal{I}} \mathcal{J}_r(u(\mathcal{I}_j))$$

and $u^* = \arg\min_{\mathcal{Q}u=0} \mathcal{J}_r(u) = u\big(\arg\min_{\mathcal{I}_j \subset \mathcal{I}} \mathcal{J}_r(u(\mathcal{I}_j))\big)$. $\qquad\qquad\square$

In fact, Proposition 4.2.1 extends to a much larger class of free-discontinuity type minimization problems; by the same reasoning as before, we arrive at the more general result:

Theorem 4.2.2. *The constrained minimization problem*

$$\begin{cases} \textit{Minimize} \quad \mathcal{J}_r^p(u) = \left[\|Tu - g\|_{\ell_2^M}^2 + \sum_{i=1}^N \min\{|u_i|^p, r^p\}\right] \\ \textit{subject to} \quad \mathcal{Q}u = 0. \end{cases} \tag{4.16}$$

has minimizers u^* for any real-valued parameter $p \ge 1$.

The Frank-Wolfe theorem, which guarantees the existence of minimizers for quadratic programs with bounded objective function, does not apply to the general case $p \ge 1$

147

where the objective function $\mathcal{J}_r^p$ is not necessarily quadratic. Nevertheless, with the following generalization for the Frank-Wolfe theorem, Theorem 4.2.2 follows directly from a similar argument as for Proposition 4.2.1.

Proposition 4.2.3. *Suppose A is an $N \times N$ positive semidefinite matrix, and suppose b and c are $N \times 1$ vectors. Suppose also that X is a nonempty convex polyhedral subset of $\mathbb{R}^N$. The convex optimization problem*

$$
\begin{cases}
minimize & u^t A u + b^t u + \sum_{1 \leq j \leq N} c_j |u_j|^p \\
subject\ to & u \in X.
\end{cases}
\tag{4.17}
$$

admits minimizers for any real parameter $p \geq 1$, as long as the objective function is bounded from below.

For ease of presentation, we reserve the proof of Proposition 4.2.3 to the Appendix.

From the proof of Theorem 4.2.2, one could in principle obtain a minimizer for $\mathcal{J}_r^p$ by computing a minimizer $u(\mathcal{I}_j)$ for each subset $\mathcal{I}_j \subset \mathcal{I}$ using a quadratic program solver [7], and then minimizing $\mathcal{J}_r^p$ over the finite set of points $\{u(\mathcal{I}_j)\}$. Unfortunately, this algorithm is computationally infeasible as the number of subsets of the index set $\{1, 2, ..., N\}$ grows exponentially with the dimension N of the underlying space.

Indeed, the minimization problem (4.16) is NP hard, as the known NP-complete problem SUBSET-SUM can be reduced to this problem. A complete discussion about the NP-hardness of (4.16) can be found in [1] .

4.3 An iterative thresholding algorithm for 1-D free-discontinuity inverse problems

4.3.1 Overview of the algorithm

In this section, we introduce an algorithm that is guaranteed to converge to a local minimizer of the real-valued functional $\mathcal{J}_r^p : \ell_2(\mathcal{I}) \to \mathbb{R}$ having the form

$$\mathcal{J}_r^p(u) = \|Tu - g\|_{\ell_2(\mathcal{K})}^2 + \sum_{i \in \mathcal{I}} \min\{|u_i|^p, r^p\}, \qquad (4.18)$$

subject to the conditions:

- $\mathcal{I}$ and $\mathcal{K}$ are countable sets of indices, and $T : \ell_2(\mathcal{I}) \to \ell_2(\mathcal{K})$ is a bounded linear operator, which is in the following identified with its matrix associated to the canonical basis;

- the operator T has spectral norm $\|T\| < 1$. Note that this requirement is easily met by an appropriate scaling for the functional, i.e., we may have to consider instead

$$\mathcal{J}_r^p(u) = \gamma\|Tu - g\|_{\ell_2(\mathcal{K})}^2 + \gamma \sum_{i \in \mathcal{I}} \min\{|u_i|^p, r^p\}, \quad \gamma \leq 1.$$

This modification leads to minor changes in the analysis that follows (see also Subsection 4.7.2), and throughout this paper we assume, without loss of generality, that $\gamma = 1$;

- the parameter p is in the range $1 \leq p \leq 2$. In case the index set $\mathcal{I}$ is *finite*, only the restriction $p \geq 1$ is necessary.

We note that the scaled 1D discrete Mumford-Shah functional $\frac{1}{h}\mathcal{J}_{1/\sqrt{h}}$ is clearly a functional of the form (4.18) having $r = 1/\sqrt{h}$, index set $\mathcal{I} = \{1, \ldots, \lfloor r^2 \rfloor\}$, parameter $p = 2$, and operator $T = D^{\dagger}_{1/r^2} : \mathbb{R}^{\lfloor r^2 \rfloor - 1} \to \mathbb{R}^{\lfloor r^2 \rfloor}$. As shown in the Appendix, the operators $D^{\dagger}_{1/r^2}$ satisfy the uniform bound $\|D^{\dagger}_{1/r^2}\| \leq 1/2$, independent of dimension, so a scaling factor is not needed in this case.

In the following, we will not minimize $\mathcal{J}_r^p$ directly. Instead, we propose a *majorization-minimization* algorithm for finding solutions to $\mathcal{J}_r^p$, motivated by the recent application of such algorithms for minimizing energy functionals arising in sparse signal recovery and image denoising [10, 30]. More precisely, consider the following *surrogate* objective function,

$$\mathcal{J}_r^{p,surr}(u,a) := \mathcal{J}_r^p(u) - \|Tu - Ta\|_{\ell_2(\mathcal{K})}^2 + \|u - a\|_{\ell_2(\mathcal{I})}^2. \quad u, a \in \ell_2(\mathcal{I}). \quad (4.19)$$

The surrogate functional $\mathcal{J}_r^{p,surr}$ satisfies $\mathcal{J}_r^{p,surr}(u,a) \geq \mathcal{J}_r^p(u)$ everywhere, with equality if and only if $u = a$, and is such that the sequence

$$u^{n+1} = \arg\min_u \mathcal{J}_r^{p,surr}(u, u^n) \quad (4.20)$$

obtained by successive minimizations of $\mathcal{J}_r^{p,surr}(u,a)$ in u for fixed a results in a nonincreasing sequence of the original functional $\mathcal{J}_r^p(u^n)$ (see Lemmas 4.3.1 and 4.3.2). We will study the implementation and the convergence properties of the

150

iteration (4.20) as follows:

- in Section 3.2, we review the standard properties of majorization-minimization iterations,

- in Section 3.3, we explicitly compute u-global minimizers of the surrogate functional $\mathcal{J}_r^{p,surr}(u,a)$, for a fixed;

- in Section 4.1 we discuss connections between the resulting thresholding functions and thresholding functions used in compressed sensing,

- in Sections 5.1, 5.2, and 5.3, we show that the sequence $(u^n)_{n\in\mathbb{N}}$ defined by (4.20) will converge to a stationary value $\bar{u} = \arg\min_u \mathcal{J}_r^{p,surr}(u,\bar{u})$, starting from any initial value u^0 for which $\mathcal{J}_r^p(u^0) < \infty$,

- in Section 5.4, we show that such stationary values $\bar{u}$ are also local minimizers of the original functional $\mathcal{J}_r^p$ that satisfy a certain fixed point condition, and

- in Section 5.5, it is shown that any global minimizer of $\mathcal{J}_r^p$ is among the set of possible fixed points $\bar{u}$ of the iteration (4.20).

By means of the thresholding algorithm, we also show that global minimizers of the functional $\mathcal{J}_r^p$ are isolated, and moreover possess a certain segmentation property that is also shared by fixed points of the algorithm.

4.3.2 Preliminary lemmas

The lemmas in this section are standard when using surrogate functionals (see [30] and [10]), and concern general real-valued surrogate functionals of the form

$$\mathcal{F}^{surr}(u, a) = \mathcal{F}(u) - \|Tu - Ta\|^2_{\ell_2(\mathcal{K})} + \|u - a\|^2_{\ell_2(\mathcal{I})}. \tag{4.21}$$

The lemmas in this section hold independent of the specific form of the functional $\mathcal{F} : \ell_2(\mathcal{I}) \to \mathbb{R}^+$, but do rely on the restriction that $\|T\| < 1$.

Lemma 4.3.1. *If the real-valued functionals $\mathcal{F}(u)$ and $\mathcal{F}^{surr}(u, a)$ satisfy the relation (4.21) and the sequence $(u^n)_{n \in \mathbb{N}}$ defined by $u^{n+1} = \arg\min_{u \in \ell_2(\mathcal{I})} \mathcal{F}^{surr}(u, u^n)$ is initialized in such a way that $\mathcal{F}(u^0) < \infty$, then the sequences $\mathcal{F}(u^n)$ and $\mathcal{F}^{surr}(u^{n+1}, u^n)$ are non-increasing as long as $\|T\| < 1$.*

Proof. Since $\|T\| < 1$, also $\|T^*T\| < 1$, and so the operator $L = \sqrt{I - T^*T}$ is a well-defined positive operator whose spectrum is contained within a closed interval $[c, 1]$ that is bounded away from zero $c > 0$. We can then rewrite $\mathcal{F}^{surr}(u^{n+1}u^n)$ as

$\mathcal{F}^{surr}(u^{n+1}, u^n) = \mathcal{F}(u^{n+1}) + \|L(u^{n+1} - u^n)\|^2_{\ell_2(\mathcal{I})}$, from which it follows that

$$
\begin{aligned}
\mathcal{F}(u^{n+1}) &\leq \mathcal{F}(u^{n+1}) + \|L(u^{n+1} - u^n)\|^2_{\ell_2(\mathcal{I})} \\
&= \mathcal{F}^{surr}(u^{n+1}, u^n) \\
&\leq \mathcal{F}^{surr}(u^n, u^n) \\
&= \mathcal{F}(u^n) \\
&\leq \mathcal{F}(u^n) + \|L(u^n - u^{n-1})\|^2_{\ell_2(\mathcal{I})} \\
&= \mathcal{F}^{surr}(u^n, u^{n-1}), \quad\quad\quad\quad\quad\quad (4.22)
\end{aligned}
$$

where the second inequality follows from u^{n+1} being a minimizer of $\mathcal{F}^{surr}(u, u^n)$. $\square$

From Lemma 4.3.1 we obtain the following corollary:

Lemma 4.3.2. *As long as the conditions of Lemma 4.3.1 are satisfied, one can choose $N \in \mathbb{N}$ sufficiently large such that for all $n \geq N$, $\|u^{n+1} - u^n\|_{\ell_2(\mathcal{I})} \leq \epsilon$, i.e.,*

$$
\lim_{n \to \infty} \|u^{n+1} - u^n\|_{\ell_2(\mathcal{I})} = 0.
$$

Proof. From Lemma 4.3.1, it follows that $\mathcal{F}(u^n) \geq 0$ is a nonincreasing sequence, therefore it converges, and $\mathcal{F}(u^n) - \mathcal{F}(u^{n+1}) \to 0$ for $n \to \infty$. The lemma follows from (4.22), and the estimates

$$
\mathcal{F}(u^n) - \mathcal{F}(u^{n+1}) \geq \|L(u^{n+1} - u^n)\|^2_{\ell_2(\mathcal{I})} \geq (1 - \|T\|^2)\|u^{n+1} - u^n\|^2_{\ell_2(\mathcal{I})}.
$$

153

□

4.3.3 The surrogate functional $\mathcal{J}_r^{p,surr}$, its explicit minimization, and a new thresholding operator

It is not immediately clear that the surrogate functional $\mathcal{J}_r^{p,surr}$ in (4.19) is any easier to manage than its parent functional $\mathcal{J}_r^p$. However, expanding the squared terms on the right hand side of (4.19), $\mathcal{J}_r^{p,surr}(u,a)$ can be equivalently expressed as

$$
\begin{aligned}
\mathcal{J}_r^{p,surr}(u,a) &= \|u-(I-T^*T)a+T^*g\|_{\ell_2(\mathcal{I})}^2 + \sum_{i\in\mathcal{I}}\min\{|u_i|^p, r^p\} + C \\
&= \sum_{i\in\mathcal{I}}\left[(u_i-[a-T^*Ta+T^*g]_i)^2 + \min\{|u_i|^p, r^p\}\right] + C,
\end{aligned}
$$

where the term $C = C(T,a,g)$ depends only on T, a and g. Indeed, unlike the original functional $\mathcal{J}_r^p$, the surrogate functional $\mathcal{J}_r^{p,surr}$ *decouples* in the variables u_i, due to the cancellation of terms involving $\|Tu\|_{\ell_2}^2$. Because of this decoupling, global u-minimizers of $\mathcal{J}_r^{p,surr}(u,a)$, for a fixed, can be computed *component-wise* according to

$$
\bar{u}_i = \arg\min_{t\in\mathbb{R}}\left[(t-[a-T^*Ta+T^*g]_i)^2 + \min\{|t|^p, r^p\}\right], \quad i\in\mathcal{I}. \tag{4.23}
$$

One can solve (4.23) explicitly when e.g. $p=2$, $p=3/2$, and $p=1$; in the general case $p\geq 1$, we have the following result:

Proposition 4.3.3 (Minimizers of $\mathcal{J}_r^{p,surr}(u,a)$ for a fixed). .

  1. *If $p>1$, the minimization problem $\bar{u} = \arg\min_{u\in\ell_2(\mathcal{I})}\mathcal{J}_r^{p,surr}(u,a)$ can be solved*

component-wise by

$$\bar{u}_i = H_{(p,r)}([a - T^*Ta + T^*g]_i), \quad i \in \mathcal{I}, \tag{4.24}$$

where $H_{(p,r)} : \mathbb{R} \to \mathbb{R}$ *is the 'thresholding function',*

$$H_{(p,r)}(\lambda) = \begin{cases} F_p^{-1}(\lambda), & |\lambda| \le \lambda'(r,p) \\ \lambda, & |\lambda| > \lambda'(r,p). \end{cases} \tag{4.25}$$

Here, $F_p^{-1}(\lambda)$ *is the inverse of the function* $F_p(t) = t + \frac{p}{2} \operatorname{sgn} t |t|^{p-1}$, *and* $\lambda' := \lambda'(r,p) \in (r, r + \frac{p}{2}r^{p-1})$ *is the unique positive value at which*

$$(F_p^{-1}(\lambda') - \lambda')^2 + |F_p^{-1}(\lambda')|^p = r^p. \tag{4.26}$$

2. *When* $p = 1$, *the general form* (4.24) *still holds, but we have to consider two cases:*

   (a) *If* $r > 1/4$, *the thresholding function* $H_{(1,r)} : \mathbb{R} \to \mathbb{R}$ *satisfies*

$$H_{(1,r)}(\lambda) = \begin{cases} 0, & |\lambda| \le 1/2 \\ (|\lambda| - 1/2) \operatorname{sgn} \lambda, & 1/2 < |\lambda| \le r + 1/4 = \lambda'(r,1) \\ \lambda, & |\lambda| > r + 1/4 \end{cases} \tag{4.27}$$

(b) If, on the other hand, $r \leq 1/4$, the function $H_{(1,r)}$ satisfies

$$H_{(1,r)}(\lambda) = \begin{cases} 0, & |\lambda| \leq \sqrt{r} = \lambda'(r,1) \\ \lambda, & |\lambda| > \sqrt{r} \end{cases} \qquad (4.28)$$

In all cases, the function $H_{(p,r)}$ is continuous except at $\lambda'(r,p)$, where $H_{(p,r)}$ has a jump-discontinuity of size $\delta(r,p) = |\lambda' - H_{(p,r)}(\lambda')| > 0$ if $r > 0$. In particular, it holds that $\lambda'(r,p) > r$ while $H_{(p,r)}(\lambda') < r$.

We leave the proof of Proposition 4.3.3 to the Appendix.

Remark 4.3.4. In the particular case $p = 2$ corresponding to classical Mumford-Shah regularization (4.13), the thresholding function $H_{(2,r)} : \mathbb{R} \to \mathbb{R}$ has a particularly simple explicit form:

$$H_{(2,r)}(\lambda) = \begin{cases} \lambda/2, & |\lambda| \leq \sqrt{2}r \\ \lambda, & |\lambda| > \sqrt{2}r \end{cases} \qquad (4.29)$$

In addition to $H_{(2,r)}$ and $H_{(1,r)}$, the thresholding operator $H_{(3/2,r)}(\lambda)$ corresponding to $p = 3/2$ can also be computed explicitly, by solving for the positive root of a suitable polynomial of third degree. In Figure 4.2 below, we plot $H_{(2,1)}, H_{(3/2,1)}$, and $H_{(1,1)}$ with parameter $r = 1$. For general noninteger values of p, $H_{(p,r)}$ cannot be solved in closed form. However, recall the following general properties of $H_{(p,r)}$:

- $H_{(p,r)}$ is an odd function,

- $H_{(p,r)}(0) = 0$, and

156

- $H_{(p,r)}(\lambda) = \lambda$ once $|\lambda| > r + \frac{p}{2}r^{p-1}$.

In fact, we can effectively *pre*compute $H_{(p,r)}$ by numerically solving for the value of $H_{(p,r)}(\lambda_j)$ on a discrete set $\{\lambda_j\}$ of points $\lambda_j \in (0, \frac{p}{2}r^{p-1} + r]$. At λ_j, one just needs to solve the real equation

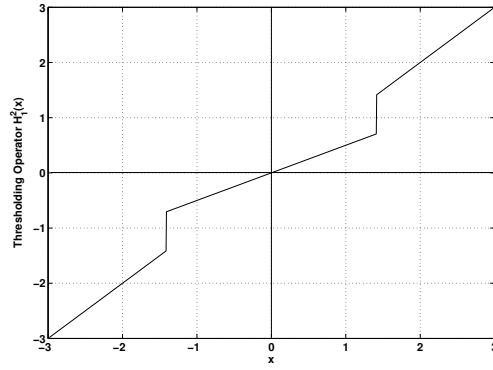$$h_j + \frac{p}{2}h_j^{p-1} - \lambda_j = 0 \tag{4.30}$$

which can be computed effortlessly via a root-finding procedure such as Newton's method: while h_j satisfies $(h_j - \lambda_j)^2 + (h_j)^p \leq r^p$, set $H_{(p,r)}(\lambda_j) = h_j$; once this constraint is violated, set $H_{(p,r)}(\lambda_j) = \lambda_j$.
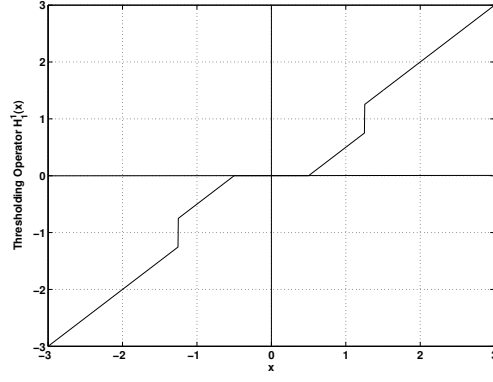
(a)
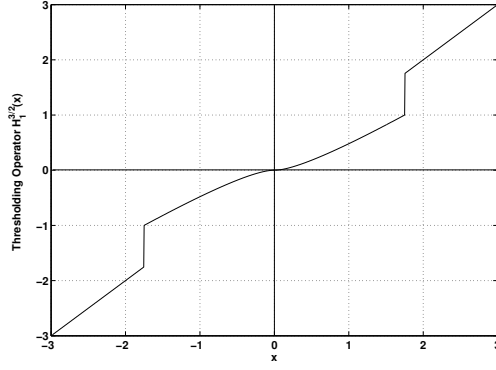
(b)

(c)

Figure 4.2: The discontinuous thresholding functions $H_{(1,1)}$, $H_{(3/2,1)}$, and $H_{(2,1)}$, with parameters $p = 1, 3/2$, and 2, respectively, and $r = 1$.

4.4 A connection to compressive sensing

When $p = 1$ and $r \leq 1/4$, we know from Theorem 4.3.3 that the iterative algorithm

$$u^{n+1} = \arg\min_u \mathcal{J}_r^{p,surr}(u, u^n) \qquad (4.31)$$

reduces to the component-wise thresholding

$$u_i^{n+1} = H_{\sqrt{r}}([u^n - T^*Tu^n + T^*g]_i), \qquad (4.32)$$

where

$$H_\gamma(\lambda) = \begin{cases} 0, & |\lambda| \leq \gamma \\ \lambda, & |\lambda| > \gamma. \end{cases} \qquad (4.33)$$

This thresholding function $H_\gamma : \mathbb{R} \to \mathbb{R}$ is referred to as *hard-thresholding* in the area of sparse recovery, and the iteration (4.32) generated by successive applications of hard thresholding has been previously studied [10]. In particular, the iteration (4.32) was shown in [10] to correspond to successive minimization in u for fixed a of the surrogate functional $\mathcal{F}_r^{0,surr}(u, a)$ corresponding to the ℓ_0 regularized functional,

$$\mathcal{F}_r^0(u) = \|Tu - g\|_{\ell_2(\mathcal{K})}^2 + r\|u\|_{\ell_0(\mathcal{I})}. \qquad (4.34)$$

Here, the ℓ_0 quasi-norm $\|u\|_{\ell_0(\mathcal{I})} := \sum_{i\in\mathcal{I}} |u_i|_0$ is defined component-wise by

$$|u_i|_0 = \begin{cases} 0, & \text{if } u_i = 0 \\ 1, & \text{otherwise} \end{cases}$$

It is not a coincidence that hard thresholding comes out from applying itera-tive thresholding to both the ℓ_0 regularized problem (4.34), and free-discontinuity problem,

$$\mathcal{J}_r^1(u) = \|Tu - g\|_{\ell_2(\mathcal{K})}^2 + \sum_{i \in \mathcal{I}} \min\{|u_i|, r\}$$

Indeed, consider the more general class of free-discontinuity-type functionals,

$$\mathcal{J}_{(r,a)}^1(u) = \|Tu - g\|_{\ell_2(\mathcal{K})}^2 + a \sum_{i \in \mathcal{I}} \min\{|u_i|, r\}, \tag{4.35}$$

which have an additional degree of freedom in the scaling parameter a that is present in the discrete MS functional (4.5), but which was omitted in the previous section to ease computations.

The free-discontinuity functional $\mathcal{J}_{(r,a)}^1$ is related to the ℓ_0 regularized functional F_γ^0 as follows:

Proposition 4.4.1. *For any sequence (r_n) satisfying $r_n \to 0$, the functional sequence $\mathcal{J}_{(\gamma/r_n, r_n)}^1$ Γ-converges to $\mathcal{F}_\gamma^0$ in the metric of ℓ_2^N.*

Proof. We have to verify the conditions for gamma convergence as established in Definition(4.1.1). Note that we can write

$$\mathcal{J}_{(\gamma/r_n, r_n)}^1 = \|Tu - g\|_{\ell_2(\mathcal{K})}^2 + \gamma \sum_{j \in \mathcal{I}} \min\{|u_j|/r_n, 1\},$$

so that Γ convergence of the full functional sequence $\mathcal{J}_{(\gamma/r_n, r_n)}^1$ is established if we can show Γ convergence of the scalar function $\min\{|x|/r_n, 1\}$ to $|x|_0$.

160

- Suppose that $x_n \to x$. Our aim is to show that $\liminf_{r_n \to 0} \min\{|x^n|/r_n, 1\} \geq |x|_0$:

  - If $|x| \geq \epsilon > 0$, then $|x^n| \geq \epsilon/2$ for n sufficiently large, and $\lim_{r_n \to 0} \min\{|x^n|/r_n, 1\} = 1$.

  - If on the other hand $x = 0$, then $\min\{|x^n|/r_n, 1\} \geq \min\{|x|/r_n, 1\} = 0$.

- On the other hand, it is easily seen that $\lim_{r_n \to 0} \min\{|x|/r_n, 1\} = |x|_0$ for any $x \in \mathbb{R}$, since

$$
\lim_{r_n \to 0} \min\{|x|/r_n, 1\} = \begin{cases} 1, & |x| > 0, \\ 0, & \text{else.} \end{cases}
$$

The desired result is established, according to (4.1.1). □

Remark 4.4.2. We showed that $\mathcal{J}^1_{(\gamma/r_n, r_n)}$ Γ converges to $\mathcal{F}^0_\gamma$ for sequences $r_n \to 0$. More generally, $\mathcal{J}^p_{(\gamma/(r_n)^p, r_n)}$ Γ converges to $\mathcal{F}^0_\gamma$ for any $p \in [1, \infty)$. It is not hard to show that the thresholding functions $H_{(p, \gamma/(r_n)^p, r_n)}$ corresponding to $\mathcal{J}^p_{(\gamma/(r_n)^p, r_n)}$, which can be derived following the proof of Proposition 4.3.3, converge to the hard thresholding function $H_{\sqrt{\gamma}}$ corresponding to $\mathcal{F}^0_\gamma$.

To ease presentation, we again take $a = 1$ in the sequel, although all of the following analysis extends to the case where a is a free parameter. Because a convergence analysis of the iteration (4.32) corresponding to hard thresholding has been studied already [10], we omit the case $p = 1$ and $r \leq 1/4$ in the sequel.

4.5 Convergence of the iterative thresholding algorithm (4.20)

4.5.1 Fixation of the discontinuity set

We prove now that the sequence $(u^n)_{n \in \mathbb{N}}$ defined by

$$u^{n+1} = \arg\min_{u} \mathcal{J}_r^{p,surr}(u, u^n) \tag{4.36}$$

or equivalently, according to Proposition 4.3.3, component-wise by

$$u_i^{n+1} = H_{(p,r)}([u^n - T^*Tu^n + T^*g]_i), \quad i \in \mathcal{I}, \tag{4.37}$$

will converge, granted that $p \geq 1$ and $\|T\| < 1$. To ease notation, we define the operator $\mathbb{H} : \ell_2(\mathcal{I}) \to \ell_2(\mathcal{I})$ by its component-wise action,

$$[\mathbb{H}(u)]_i := H_{(p,r)}([u - T^*Tu + T^*g]_i); \tag{4.38}$$

so that the iteration (4.37) can be written more concisely in operator notation as

$$u^{n+1} = \mathbb{H}(u^n). \tag{4.39}$$

We omit the dependence of $\mathbb{H}$ on the parameters p, r, and the function g for continuity of presentation. At the core of the convergence proof is the fact that the 'discontinuity set', indicated below by $\mathcal{I}_1^n$, of u^n must eventually fix during the iteration (4.37), at

162

which point the 'free-discontinuity' problem is transformed into a simpler 'fixed-discontinuity' problem.

Lemma 4.5.1 (Fixation of the index set $\mathcal{I}_1$). *Fix $p \geq 1, r \in \mathbb{R}^+$, and $g \in \ell_2(\mathcal{K})$. Consider the iteration*

$$u^{n+1} = \mathbb{H}(u^n) \tag{4.40}$$

and the time-dependent partition of the index set $\mathcal{I}$ into 'small' set

$$\mathcal{I}_0^n \;=\; \{i \in \mathcal{I} : |u_i^n| \leq \lambda'(r, p)\} \tag{4.41}$$

and 'large' set

$$\mathcal{I}_1^n \;=\; \{i \in \mathcal{I} : |u_i^n| > \lambda'(r, p)\} \tag{4.42}$$

where $\lambda'(r, p)$ is the position of the jump discontinuity of the thresholding function, as defined in Proposition 4.3.3. For $N \in \mathbb{N}$ sufficiently large, this partition fixes during the iteration $u^{n+1} = \mathbb{H}(u^n)$; that is, there exists a set $\mathcal{I}_0$ such that for all $n \geq N$, $\mathcal{I}_0^n = \mathcal{I}_0$ and $\mathcal{I}_1^n = \mathcal{I}_1 := \mathcal{I} \setminus \mathcal{I}_0$.

Proof. By *discontinuity* of the thresholding operator $H_{(p,r)}(\lambda)$, each sequence component

$$u_i^n = H_{(p,r)}([u^{n-1} - T^* T u^{n-1} + T^* g]_i) \tag{4.43}$$

satisfies

(a) $|u_i^n| \leq \lambda'(r, p) - \delta(r, p) < \lambda'(r, p)$, if $i \in \mathcal{I}_0^n$, or

(b) $|u_i^n| > \lambda'(r,p)$, if $i \in \mathcal{I}_1^n$.

Thus, $|u_i^{n+1} - u_i^n| \geq \delta(r,p)$ if $i \in \mathcal{I}_0^{n+1} \bigcap \mathcal{I}_1^n$, or vice versa if $i \in \mathcal{I}_0^n \bigcap \mathcal{I}_1^{n+1}$. At the same time, Lemma 4.3.2 implies

$$|u_i^{n+1} - u_i^n| \leq \|u^{n+1} - u^n\|_{\ell_2(\mathcal{I})} \leq \epsilon, \tag{4.44}$$

once $n \geq N(\epsilon)$, and $\epsilon > 0$ can be taken arbitrarily small. In particular, (4.44) implies that $\mathcal{I}_0$ and $\mathcal{I}_1$ must be fixed once $n \geq N(\epsilon)$ and $\epsilon < \delta(r,p)$. $\qquad\square$

After fixation of the index set $\mathcal{I}_0 = \{i \in \mathcal{I} : |u_i^n| \leq \lambda'(r,p)\}$, $\mathbb{H}(u^n) = \mathbb{U}_{\mathcal{I}_0}(u^n)$ and $\mathbb{U}_{\mathcal{I}_0}$ is an operator having component-wise action, for $p > 1$,

$$[\mathbb{U}_{\mathcal{I}_0} u]_i = \begin{cases} F_p^{-1}([(I - T^*T)u + T^*g]_i), & \text{if } i \in \mathcal{I}_0 \\ \left((I - T^*T)u + T^*g\right)_i, & \text{if } i \in \mathcal{I}_1 \end{cases} \tag{4.45}$$

Here, as in Proposition 4.3.3, the function F_p^{-1} is the inverse of the function $F_p(t) = t + \frac{p}{2}\operatorname{sgn} t |t|^{p-1}$. Again, for ease of presentation, we omit the dependence of $\mathbb{U}_{\mathcal{I}_0}$ on the parameters p, r, and g. For $p = 1$ the description is similar, and in general, one easily verifies the equivalence

$$\mathbb{U}_{\mathcal{I}_0}(v) = \arg \min_{u \in \ell_2(\mathcal{I})} \mathcal{J}_{\mathcal{I}_0}^{p,surr}(u, v) \tag{4.46}$$

where $\mathcal{J}_{\mathcal{I}_0}^{p,surr}$ is a surrogate for the *convex* functional,

$$\mathcal{J}_{\mathcal{I}_0}^p(u) := \|Tu - g\|_{\ell_2(\mathcal{K})}^2 + \sum_{i \in \mathcal{I}_0} |u_i|^p. \tag{4.47}$$

164

That is, fixation of the index set $\mathcal{I}_0$ implies that the sequence $(u^n)_{n \in \mathbb{N}}$ has become constrained to a subset of $\ell_2(\mathcal{I})$ on which the map $\mathbb{H}$ agrees with a map $\mathbb{U}_{\mathcal{I}_0}$, associated to the convex functional $\mathcal{J}_{\mathcal{I}_0}^p$. As we will see, this implies that the nonconvex functional $\mathcal{J}_r^p$ behaves *locally* like a convex functional in neighborhoods of fixed points $u = \mathbb{H}(u)$, including the global minimizers of $\mathcal{J}_r^p$.

4.5.2 On the nonexpansiveness and convergence for T injective

Given that $\mathbb{H}(u^n) = \mathbb{U}_{\mathcal{I}_0}(u^n)$ after a finite number of iterations, we can use well-known tools from convex analysis to prove that the sequence $(u^n)_{n \in \mathbb{N}}$ converges. If the operator $T^*T : \ell_2(\mathcal{I}) \to \ell_2(\mathcal{K})$ is invertible, or, equivalently, if the operator T maps onto its range and has a trivial null space – as, for example, does the discrete pseudoinverse $D_h^\dagger$ in the 1D Mumford-Shah approximation – then the mapping $\mathbb{U}_{\mathcal{I}_0}$ has the nice property of being a contraction mapping, so that a direct application of the Banach fixed point theorem ensures exponential convergence of the sequence $(u^n)_{n \in \mathbb{N}}$ after fixation of the index sets.

Theorem 4.5.2. *Suppose $T : \ell_2(\mathcal{I}) \to \ell_2(\mathcal{K})$ maps onto $\ell_2(\mathcal{K})$ and has a trivial null space. Let $\delta > 0$ be a lower bound on the spectrum of T^*T. Then the sequence*

$$u^{n+1} = \mathbb{H}(u^n), \tag{4.48}$$

as defined in (4.38), is guaranteed to converge in norm. In particular, after a finite

number of iterations $N \in \mathbb{N}$, this mapping takes the form

$$u^{N+m} = \mathbb{U}_{\mathcal{I}_0}^m(u^N), \quad m \in \mathbb{N} \setminus \{0\}, \tag{4.49}$$

and the sequence $(u^n)_{n \in \mathbb{N}}$ converges to the unique fixed point $\bar{u}$ of the map $\mathbb{U}_{\mathcal{I}_0}$. Moreover, after fixation of of the index set $\mathcal{I}_0$, the rate of convergence becomes exponential:

$$\|u^{N+m} - \bar{u}\|_{\ell_2(\mathcal{I})} = \|\mathbb{U}_{\mathcal{I}_0}^m(u^N) - \mathbb{U}_{\mathcal{I}_0}^m(\bar{u})\|_{\ell_2(\mathcal{I})} \leq (1-\delta)^m \|u^N - \bar{u}\|_{\ell_2(\mathcal{I})}, \quad m \in \mathbb{N} \setminus \{0\}. \tag{4.50}$$

The proof of Theorem 4.5.2 is deferred to the Appendix.

4.5.3 Convergence for general operators T

Unfortunately, if T^*T is not invertible (that is, if $\delta = 0$ belongs to its nonnegative spectrum), then the map $\mathbb{U}_{\mathcal{I}_0}$ is not necessarily a contraction, and we can no longer apply the Banach fixed point theorem to prove convergence of the sequence $(u^n)_{n \in \mathbb{N}}$. However, as long as $\|T\| < 1$, we observe by following the proof of Theorem (4.5.2) that $\mathbb{U}_{\mathcal{I}_0}$ is still *non-expansive*, meaning that for all $v, v' \in \ell_2(\mathcal{I})$, $\|\mathbb{U}_{\mathcal{I}_0}(v) - \mathbb{U}_{\mathcal{I}_0}(v')\|_{\ell_2(\mathcal{I})} \leq \|v - v'\|_{\ell_2(\mathcal{I})}$. The following Opial's theorem [60], here reported adjusted to our notations and context, gives sufficient conditions under which non-expansive maps admit convergent successive iterations:

Theorem 4.5.3 (Opial's Theorem). *Let the mapping $\mathbb{A}$ from $\ell_2(\mathcal{I})$ to $\ell_2(\mathcal{I})$ satisfy the following conditions:*

166

1. $\mathbb{A}$ *is asymptotically regular: for all* $v \in \ell_2(\mathcal{I})$, $\|\mathbb{A}^{n+1}(v) - \mathbb{A}^n(v)\|_{\ell_2(\mathcal{I})} \to 0$ *for* $n \to \infty$;

2. $\mathbb{A}$ *is non-expansive: for all* $v, v' \in \ell_2(\mathcal{I})$, $\|\mathbb{A}(v) - \mathbb{A}(v')\|_{\ell_2(\mathcal{I})} \leq \|v - v'\|_{\ell_2(\mathcal{I})}$;

3. *the set* $\mathrm{Fix}(\mathbb{A})$ *of the fixed points of* $\mathbb{A}$ *in* $\ell_2(\mathcal{I})$ *is not empty.*

Then, for all $v \in \ell_2(\mathcal{I})$, *the sequence* $(\mathbb{A}^n(v))_{n \in \mathbb{N}}$ *converges weakly to a fixed point in* $\mathrm{Fix}(\mathbb{A})$.

In fact, we already know that $\mathbb{U}_{\mathcal{I}_0}$ is asymptotically regular, in addition to being nonexpansive - this follows by application of Lemma 4.3.1 and Lemma 4.3.2 to the functional $\mathcal{J}_{\mathcal{I}_0}^p$. Thus, in order to apply Opial's theorem, it remains only to show that $\mathbb{U}_{\mathcal{I}_0}$ has a fixed point; that is, that there exists a point $\bar{u} \in \ell_2(\mathcal{I})$ for which

$$\bar{u} = \mathbb{U}_{\mathcal{I}_0}(\bar{u}).$$

In more detail, we must prove the existence of a vector $\bar{u} \in \ell_2(\mathcal{I})$ satisfying

$$\bar{u}_i \;=\; \begin{cases} F_p^{-1}([(I - T^*T)\bar{u} + T^*g]_i), & \text{if } i \in \mathcal{I}_0 \\ \left((I - T^*T)\bar{u} + T^*g\right)_i, & \text{if } i \in \mathcal{I}_1 \end{cases} \tag{4.51}$$

The following lemma gives a simple yet useful characterization of points satisfying the fixed point relation (4.51):

Lemma 4.5.4. *Suppose* $p > 1$. *A vector* $\bar{u} \in \ell_2(\mathcal{I})$ *satisfies the fixed point relation*

167

$\bar{u} = \mathbb{U}_{\mathcal{I}_0}(\bar{u})$ *if and only if*

$$
\left[T^*(g - T\bar{u})\right]_i = \begin{cases} 0, & i \in \mathcal{I}_1 \\[2mm] F_p(\bar{u}_i) - \bar{u}_i, & i \in \mathcal{I}_0, \end{cases} \tag{4.52}
$$

Alternatively, if $p = 1$ and $r \geq 1/4$, $\bar{u} = \mathbb{U}_{\mathcal{I}_0}(\bar{u})$ is satisfied if and only if

$$
\begin{cases} \left[T^*(g - T\bar{u})\right]_i \in [-1/2, 1/2], & i \in \mathcal{I}_0^a, \\[2mm] \left[T^*(g - T\bar{u})\right]_i = 1/2 \operatorname{sgn} \bar{u}_i, & i \in \mathcal{I}_0^b, \\[2mm] \left[T^*(g - T\bar{u})\right]_i = 0, & i \in \mathcal{I}_1, \end{cases} \tag{4.53}
$$

where in (4.53), the index set $\mathcal{I}_0$ is split into

- $\mathcal{I}_0^a = \{i \in \mathcal{I}_0 : |\bar{u}_i| \leq 1/2\}$, *and*

- $\mathcal{I}_0^b = \{i \in \mathcal{I}_0 : 1/2 < |\bar{u}_i| \leq r + 1/4\}$.

Again, recall the notation $F_p(t) = t + \frac{p}{2} \operatorname{sgn} t |t|^{p-1}$, and observe that the fixed point relation (4.52) has a very simple expression when $p = 2$. The proof of Lemma 4.5.4 is given in the Appendix.

The fixed point characterization of Lemma 4.5.4 will be crucial in the following theorem that ensures the existence of a fixed point $\bar{u} = \mathbb{U}_{\mathcal{I}_0}(\bar{u})$. We remind the reader that until now, all of the results of Section 1.3 remain valid in the infinite-dimensional setting $|\mathcal{I}| = \infty$. From this point on, however, certain results will only hold in finite dimensions; for clarity, we will account each such situation explicitly.

Proposition 4.5.5. *In finite dimensions* $|\mathcal{I}| < \infty$, *then there exist (global) minimizers of the convex functional,*

$$\mathcal{J}^p_{\mathcal{I}_0}(u) = \|Tu - g\|^2_{\ell_2(\mathcal{K})} + \sum_{i \in \mathcal{I}_0} |u_i|^p, \tag{4.54}$$

for all $p \geq 1$, *and any minimizer* $\bar{u}$ *of* $\mathcal{J}^p_{\mathcal{I}_0}$ *satisfies the fixed point relation* $\bar{u} = \mathbb{U}_{\mathcal{I}_0}(\bar{u})$. *Restricted to the range* $1 \leq p \leq 2$, *the statement is true also in the limit* $|\mathcal{I}| = \infty$.

Proof. In the finite-dimensional setting, minimizers necessarily exist for all $p \geq 1$ according to Proposition 4.2.3. We now consider the general case. Consider the unique decomposition $u = u_0 + u_1$ into a vector u_0 supported on $\mathcal{I}_0$ and another u_1 supported on $\mathcal{I}_1$, i.e., the vectors $u_0 \in \ell_2^{\mathcal{I}_0}(\mathcal{I}) := \{u \in \ell_2(\mathcal{I}) : u_i = 0, \quad i \in \mathcal{I}_1\}$ and $u_1 \in \ell_2^{\mathcal{I}_1}(\mathcal{I}) := \{u \in \ell_2(\mathcal{I}) : u_i = 0, \quad i \in \mathcal{I}_0\}$. Let $\mathcal{P} : u \to u_1$ and $\mathcal{P}^\perp = \mathcal{I} - \mathcal{P} : u \to u_0$ denote the orthogonal projections onto the subspaces $\ell_2^{\mathcal{I}_1}(\mathcal{I})$ and $\ell_2^{\mathcal{I}_0}(\mathcal{I})$, respectively. Consider the operators $T_0 = T\mathcal{P}^\perp$ and $T_1 = T\mathcal{P}$; note that clearly $T = T_0 + T_1$ is satisfied. The functional (4.54) can be re-written with this decomposition according to

$$\mathcal{J}^p_{\mathcal{I}_0}(u_0 + u_1) = \|T_0 u_0 + T_1 u_1 - g\|^2_{\ell_2(\mathcal{K})} + \|u_0\|^p_{\ell_p^{\mathcal{I}_0}(\mathcal{I})} \tag{4.55}$$

where $\|z\|_{\ell_p^{\mathcal{I}_0}(\mathcal{I})} := \left(\sum_{i \in \mathcal{I}_0} |z_i|^p\right)^{1/p}$ is the ℓ_p-norm on vectors supported on $\mathcal{I}_0$.

Let $\mathcal{P}_1$ be the orthogonal projection onto the range of T_1 in $\ell_2(\mathcal{K})$ (not to be confused with $\mathcal{P}$, which operates on the space $\ell_2(\mathcal{I})$) and let $\mathcal{P}_1^\perp = \mathcal{I} - \mathcal{P}_1$ be the orthogonal projection in $\ell_2(\mathcal{K})$ onto the orthogonal complement of the range of T_1. Then, fixing

$u_0 \in \ell_2^{\mathcal{I}_1}(\mathcal{I})$, the vector $\mathcal{P}_1(g - T_0 u_0) \in \text{range}(T_1) \subset \ell_2(\mathcal{K})$ is the solution to the minimization problem

$$\mathcal{P}_1(g - T_0 u_0) = \arg \min_{v \in \text{range}(T_1)} \|v - (g - T_0 u_0)\|_{\ell_2(\mathcal{K})}^2, \tag{4.56}$$

so that minimizers of the functional $\mathcal{F} : \ell_2^{\mathcal{I}_0}(\mathcal{I}) \to \mathbb{R}^+$ defined by

$$
\begin{aligned}
\mathcal{F}(v) &= \|T_0 v + \mathcal{P}_1(g - T_0 v) - g\|_{\ell_2(\mathcal{K})}^2 + \|v\|_{\ell_p^{\mathcal{I}_0}(\mathcal{I})}^p \\
&= \|K v - y\|_{\ell_2(\mathcal{K})}^2 + \|v\|_{\ell_p^{\mathcal{I}_0}(\mathcal{I})}^p
\end{aligned} \tag{4.57}
$$

with $K := \mathcal{P}_1^\perp T_0$, and $y := \mathcal{P}_1^\perp g$, will yield minimizers of $\mathcal{J}_{\mathcal{I}_0}^p$. Functionals of the form (4.57) were studied in [30]; there, it is shown that as long as $1 \leq p \leq 2$, $\mathcal{F}$ has minimizers, and any minimizer $\bar{v}$ can be characterized by the fixed point relation

$$\bar{v}_i = F_p^{-1}([(I - K^* K)\bar{v} + K^* y]_i), \quad i \in \mathcal{I}_0; \tag{4.58}$$

(recall that F_p^{-1} is the inverse of the function $F_p(t) = t + \frac{p}{2} \operatorname{sgn} t |t|^{p-1}$).

In the finite-dimensional setting $|\mathcal{I}| < \infty$, the Euler-Lagrange equations corresponding to minimizers of the convex functional $\mathcal{F}$ as in (4.57) imply the same fixed point relation (4.58) also, for all $p \geq 1$.

By Lemma 4.5.4, the characterization (4.58) is equivalent to the condition

- $p > 1$:

$$\left[K^*(y - K\bar{v})\right]_i = \frac{p}{2} \operatorname{sgn} \bar{v}_i |\bar{v}_i|^{p-1}, \tag{4.59}$$

- $p = 1$:

$$\begin{cases} \left[K^*(y - K\bar{v})\right]_i \in [-1/2, 1/2], & \text{if } |\bar{v}_i| \leq 1/2, \\ \left[K^*(y - K\bar{v})\right]_i = 1/2 \operatorname{sgn} \bar{v}_j, & \text{if } 1/2 < |\bar{v}_i| \leq r + 1/4. \end{cases} \quad, \quad i \in \mathcal{I}_0. \quad (4.60)$$

Making the identification $\bar{u}_0 = \bar{v}$ and $T_1 \bar{u}_1 = \mathcal{P}_1(g - T_0 \bar{v})$, and rewriting $K = \mathcal{P}_1^\perp T_0$, and $y = \mathcal{P}_1^\perp g$, the relations (4.59) and (4.60) imply the full fixed point characterization in Lemma 4.5.4. $\qquad\square$

Remark 4.5.6. The restriction $p \leq 2$ that is necessary for the results of this paper in the infinite dimensional setting $|\mathcal{I}| = \infty$ was only used in the proof of Theorem 4.5.5, where it comes from [30] and is needed there to prove the existence of minimizers of functionals $\mathcal{F}$ of the form (4.57). If that proof can be extended to functionals of the form (4.57) for general $p \geq 1$, then the restriction $p \leq 2$ can be dropped in the current paper. For instance, if we additionally require that T is a bounded operator from $\ell_p(\mathcal{I})$ to $\ell_2(\mathcal{I})$ for $1 \leq p < \infty$ then the existence of minimizers would be guaranteed also for $1 \leq p < \infty$ and $|\mathcal{I}| = \infty$. In this case we could consider a minimizing sequence (v^k) of $\mathcal{F}$, which is necessarily bounded in ℓ_p. Therefore, there exists a subsequence (v^{k_h}) which weakly converges in ℓ_p to a point v^*. This also implies the weak convergence of the sequence Kv^{k_h} in ℓ_2; note that $\langle Kv^{k_h}, w \rangle_{\ell_2 \times \ell_2} = \langle v^{k_h}, K^* w \rangle_{\ell_p \times \ell_{p'}}$, for $1/p + 1/p' = 1$. By Fatou's lemma we obtain $\mathcal{F}(v^*) \leq \liminf_h \mathcal{F}(v^{k_h})$ and v^* is a minimizer of $\mathcal{F}$. However, we still require that $p \geq 1$ for the proof of Proposition 4.3.3 and for the results of the next section to hold.

Combining the results from this section, we obtain:

Theorem 4.5.7. *Suppose $1 \leq p \leq 2$. Starting from any u^0 satisfying $\mathcal{J}_r^p(u^0) < \infty$, the sequence $(u^n)_{n \in \mathbb{N}}$ defined by $u^{n+1} = \mathbb{H}^n(u^0)$ as in (4.38) will converge weakly to a vector $\bar{u} \in \ell_2(\mathcal{I})$ that satisfies the fixed point condition,*

1. *$|\bar{u}_i| \geq \lambda'(r, p)$, if $i \in \mathcal{I}_1 = \{j \in \mathcal{I} : |\bar{u}_j| > r\}$*

2. *$|\bar{u}_i| \leq F_p^{-1}(\lambda'(r, p))$, for $p > 1$, if $i \in \mathcal{I}_0 = \{j \in \mathcal{I} : |\bar{u}_j| \leq r\}$, and*

3. *(a) If $p > 1$:*

$$
\left[T^*(g - T\bar{u})\right]_i =
\begin{cases}
0, & \text{if } |\bar{u}_i| \geq \lambda'(r, p) \\
F_p(\bar{u}_i) - \bar{u}_i, & \text{if } |\bar{u}_i| \leq \lambda'(r, p) - \delta(r, p)
\end{cases}
\tag{4.61}
$$

(b) If $p = 1$ and $r \geq 1/4$:

$$
\begin{cases}
\left[T^*(g - T\bar{u})\right]_i \in [-1/2, 1/2], & |\bar{u}_i| \leq 1/2 \\
\left[T^*(g - T\bar{u})\right]_i = 1/2 \operatorname{sgn} \bar{u}_i, & 1/2 < |\bar{u}_i| \leq r - 1/4. \\
\left[T^*(g - T\bar{u})\right]_i = 0, & |\bar{u}_i| > r + 1/4.
\end{cases}
\tag{4.62}
$$

If the index set $|\mathcal{I}| < \infty$ is finite dimensional, the theorem holds for all $p \geq 1$.

Proof. By Lemma 4.5.1, the map $u^{n+1} = \mathbb{H}(u^n)$ becomes equivalent to a map of the form $u^{n+1} = \mathbb{U}_{\mathcal{I}_0}(u^n)$ after a finite number of iterations $N \in \mathbb{N}$. By Lemma 4.5.1 and Proposition 4.3.3, the subset $\mathcal{I}_0 \subset \mathcal{I}$ separates $\mathcal{I}$ in the sense that, for all $n \geq N$,

- $|u_i^n| < F_p^{-1}(\lambda'(r, p))$, if $i \in \mathcal{I}_0$,

- $|u_i^n| > \lambda'(r, p)$, if $i \in \mathcal{I}_1 = \mathcal{I} \setminus \mathcal{I}_0$.

That the sequence $(u^n)_{n \in \mathbb{N}}$ converges to a fixed point of the map $\mathbb{U}_{\mathcal{I}_0}$ follows from Opial's theorem applied to the map $\mathbb{U}_{\mathcal{I}_0}$:

1. the asymptotic regularity of $\mathbb{U}_{\mathcal{I}_0}$ is a consequence of Lemmas 4.3.1 and 4.3.2;

2. the nonexpansiveness of $\mathbb{U}_{\mathcal{I}_0}$ follows from the proof of Theorem (4.5.2), and

3. Theorem 4.5.5 guarantees that the set of fixed points of $\mathbb{U}_{\mathcal{I}_0}$ in $\ell_2(\mathcal{I})$ is nonempty.

The limit $\bar{u}$ of the sequence (u^n) will satisfy the fixed point conditions of Lemma 4.5.4. Since weak convergence implies component-wise convergence, it follows for all $i \in \mathcal{I}_0$ that

$$
\begin{aligned}
|\bar{u}_i| &= \lim_{n \to \infty} |u_i^n| \\
&\leq \lambda'(r, p) - \delta(r, p)
\end{aligned}
\tag{4.63}
$$

and the respective lower bound $|u_i^n| \geq \lambda'(r, p)$ holds analogously for $i \in \mathcal{I}_1$. $\qquad \square$

4.6   On minimizers of $\mathcal{J}_r^p$

We are now in a position to explore the relationship between limit vectors $\bar{u}$ of the iterative thresholding algorithm (4.38) and minimizers of the free-discontinuity functional $\mathcal{J}_r^p$ (4.18). As a first but important result in this direction,

Theorem 4.6.1. *A point $\bar{u}$ satisfying the fixed point relation of Theorem 4.5.7 is a local minimizer of the functional $\mathcal{J}_r^p$ defined in (4.18).*

The proof of Theorem 4.6.1 is omitted at present but can be found in the Appendix. This result should not be surprising, however. Due to the separation of the entries of any fixed point $\bar{u}$, such that $\bar{u}_i < r < \bar{u}_j$ for $i \in \mathcal{I}_0$ and $j \in \mathcal{I}_1$, we have also $\mathcal{I}_0 \equiv \{i \in \mathcal{I} : |u_i| \leq r\}$ and $\mathcal{I}_1 \equiv \{j \in \mathcal{I} : |u_j| > r\}$ for all $u \in B(\bar{u}, \varepsilon(r))$, where $B(\bar{u}, \varepsilon(r))$ is a ball around an equilibrium point $\bar{u}$ of radius $\varepsilon(r) > 0$ sufficiently small. On this neighborhood $B(\bar{u}, \varepsilon(r))$ of $\bar{u}$, the functional $\mathcal{J}_r^p$ is convex. Since $\bar{u}$ is obtained as the limit of a sequence (u^n) in $B(\bar{u}, \varepsilon(r))$ for which the sequence $\mathcal{J}_r^p(u^n)$ is nonincreasing, one would expect that $\bar{u}$ minimizes $\mathcal{J}_r^p(u^n)$ within this neighborhood.

More surprising is that global minimizers of $\mathcal{J}_r^p$ are also fixed points, as shown in the following theorem. Even though the existence of such minimizers is only guaranteed in the finite-dimensional setting (see Proposition 4.2.3), the following result is not restricted as such.

Theorem 4.6.2 (Global minimizers of $\mathcal{J}_r^p$ are fixed points $\bar{u} = \mathbb{H}(\bar{u})$). *Any global minimizer u^* of $\mathcal{J}_r^p$ satisfies the fixed point condition of the map $\mathbb{H}$ that is given in Theorem 4.5.7.*

The proof of Theorem 4.6.2 is rather long and we defer it to the Appendix. We reiterate once more that on a ball $B(\bar{u}, \varepsilon(r))$ around an equilibrium point $\bar{u}$ of radius $\varepsilon(r) > 0$ sufficiently small, the functional $\mathcal{J}_r^p$ is convex; following the proof of Theorem 4.6.2, we see that $\mathcal{J}_r^p$ is in fact *strictly* convex whenever $\bar{u} = u^*$ is a global minimizer, since the restriction of T to the subspace $\ell_2^{\mathcal{I}_1}(\mathcal{I}) \subset \ell_2(\mathcal{I})$ of vectors with support in $\mathcal{I}_1$ must be an injective operator in this case. Hence a global minimizer

is necessarily an isolated minimizer, whereas we cannot ensure the same property for local minimizers if T has a nontrivial null-space; in this case, local minimizers may form continuous sets, as it is shown in the bottom-right box of Figure 4.3. We conclude the following remark.

Corollary 4.6.3. *Minimizers of $\mathcal{J}_r^p$ are isolated.*

4.7 Numerical Experiments

4.7.1 Dynamical systems, stability, and equilibria

Iterative thresholding algorithms have a natural interpretation as discrete-time dynamical systems with nonsmooth right-hand-side, and can be associated to continuous dynamical systems of the type:

$$
\begin{aligned}
\dot{u}(t) &= F(u(t), t) \\
&= \tau \left(H_{(p,r)}(u(t) + T^*(g - Tu(t))) - u(t) \right), \quad t \geq t_0, \quad \tau > 0.
\end{aligned}
$$

The study of the existence, uniqueness, stability, and long-time behavior of these ODE's is of fundamental interest in order to clarify also the stability properties of iterative thresholding algorithms. Indeed, other than soft-thresholding iterations [30], the corresponding right-hand-side is not Lipschitz continuous and can even be discontinuous, as is the case for free-discontinuity problems. In [14, 42] conditions are established for the existence, uniqueness, and continuous dependence on the initial data (at finite time) of solutions of dynamical systems with discontinuous
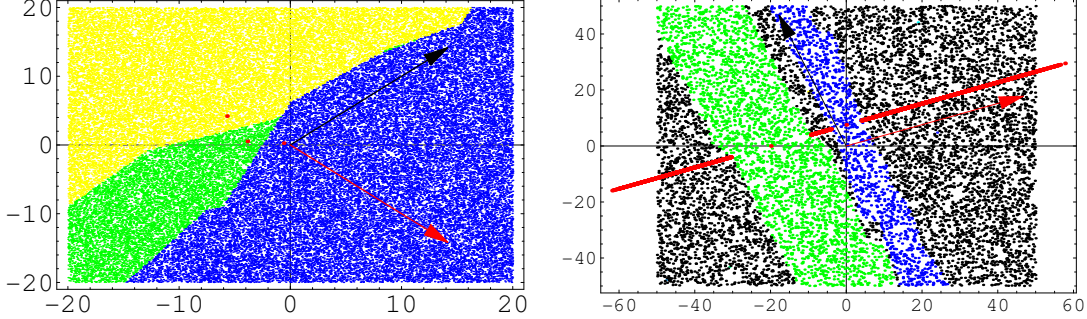
Figure 4.3: We show patterns in $\mathbb{R}^2$ formed by initial points u^0 colored according to the corresponding equilibria computed as limits of the iterative thresholding algorithm (4.37). For invertible 2×2 squared matrices T, the equilibria are isolated and the region of initial points for which (4.37) converges to a given equilibrium point do partition the space into sets which might be disconnected. Structures of the partition generated by a particular matrix T is exemplified on the left, and in the right box we show a pattern related to iterations where the 2×2 squared matrix T has nontrivial null-space. We can see again that global minimizers are isolated and correspond to the points on the axes, whereas local minimizers are continuously distributed along an affine space generated by the kernel of T. It is not difficult to show that this structure always occurs for such matrices.

right-hand-side. However, very little is known about long-time properties of such dynamical systems and about the nature of their equilibrium points.

For several continuous thresholding functions, such as the ones introduced in [30, 45, 44], one can easily show, for instance by means of Γ-convergence arguments, that equilibrium points depend continuously on the parameters of the thresholding, see, e.g., [44, Theorem 5.1]. Nevertheless, for discontinuous thresholding functions $H_{(p,r)}$ such as those studied in this paper, sudden bifurcation phenomena and instabilities do appear in general. Figure 4.3 shows that multiple equilibrium points can ex-

176

ist for these thresholding operators and their number may depend discontinuously on the thresholding shape parameters. Moreover, as established in Theorem 4.6.2, global minimizers of $\mathcal{J}_r^p$ are always stable equilibria and isolated points, while local minimizers can be unstable equilibria and form a continuous set, as shown in the bottom-right box of Figure 4.3.

4.7.2 Denoising and segmentation of 1-D signals and digital images

In this subsection, we are concerned with numerical experiments in the use of an iterative thresholding algorithm for the minimization of

$$\mathcal{J}_{r,\gamma}^2(u) := \|D_h^\dagger u - g\|_{\ell_2}^2 + \gamma \sum_{i=1}^N \min\{u_i^2, r^2\}, \tag{4.64}$$

modelling problems of denoising and segmentation.

Note that we introduced an additional regularization parameter $\gamma > 0$ which has the sole effect of modifying the thresholding function $H_{(2,r,\gamma)}$ as follows

$$H_{(2,r,\gamma)}(z) = \begin{cases} \frac{1}{1+\gamma}z, & |z| \le \frac{\sqrt{1+\gamma}r}{\sqrt{\gamma}} \\ z, & \text{otherwise.} \end{cases} \tag{4.65}$$

This thresholding function can be again easily computed by means of an argument similar to the proof of Proposition 4.3.3. In Figure **??** we show the results of applications of the iterative thresholding algorithm (4.37). In Figure 4.5 we show a comparison of the use of the thresholding $H_{(2,r,\gamma)}$ and the soft-thresholding S_γ (see
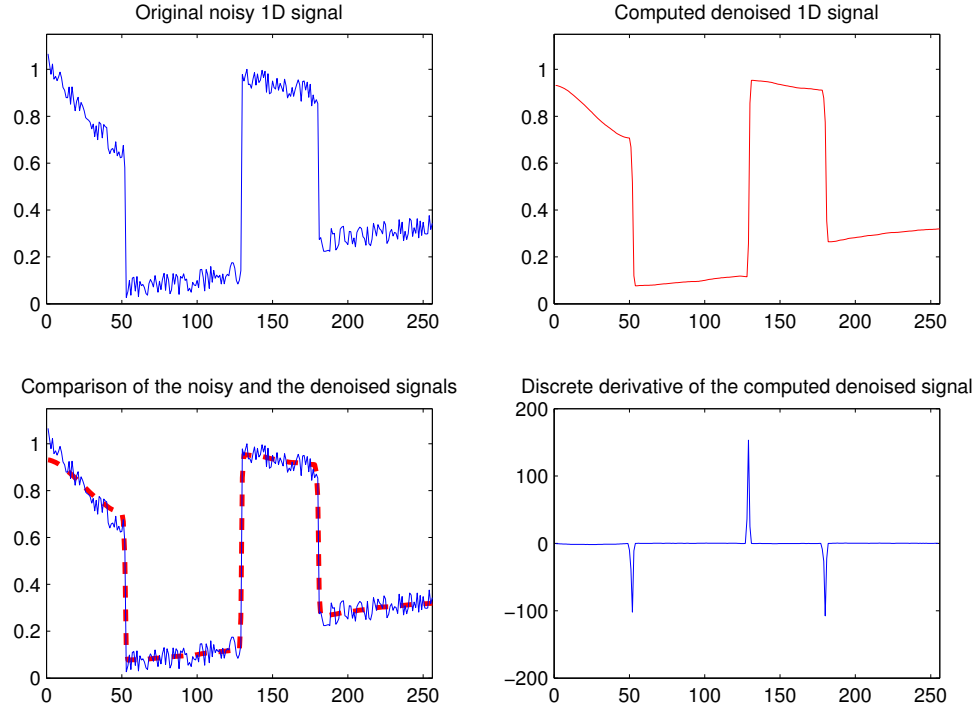
Figure 4.4: We show the application of the iterative thresholding algorithm (4.37) for the classical denoising problem of 1-D signals where $K = I$ in (4.10), and hence $T = D_h^\dagger$. The thresholding parameters used for the numerics are $r = 2.2$ and $\gamma = 0.002$.
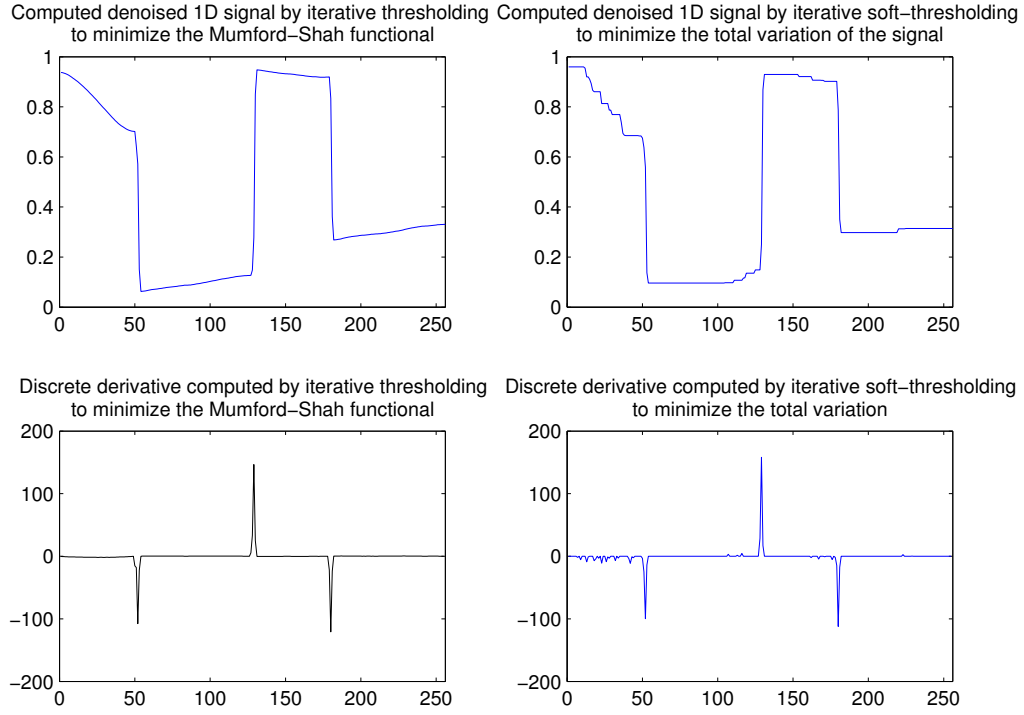
Figure 4.5: A comparison of the denoising of the signal in Figure **??** by means of the algorithm (4.37) and by iterative soft-thresholding [30] applied to discrete derivatives. We can appreciate how the algorithm (4.37) promotes piecewise smooth solutions, whereas the iterative soft-thresholding promotes the total variation minimization with the introduction of a 'staircase effect'. The thresholding parameters used for the numerics are $r = 2.2$ and $\gamma = 0.002$ for (4.65), and $\gamma = 0.002$ for the soft-thresholding (4.104).

its definition in (4.104)); the former promotes the minimization of the Mumford-Shah constraint MS and piecewise smooth solutions, whereas the latter promotes the minimization of a total variation constraint [67], which is also well-known to produce (almost) piecewise constant solutions with a perhaps unwanted 'staircase effect'; see also [23, Section 4] for details.

4.7.3 Inverse problems

As already mentioned in Subsection 4.1.4 the Mumford-Shah term $MS(u) = \int_{\Omega \setminus S_u} |\nabla u|^2 + \beta \mathcal{H}^{d-1}(S_u)$ is also used for regularizing inverse problems involving operators T which are not boundedly invertible. In this section we present a numerical experiment on the use of the algorithm (4.37) for 1D interpolation (Figure 4.6). In this case the operator T is a multiplier by a characteristic function of a subdomain, i.e., $Tu := \chi_D \cdot u$, for $D \subset \Omega$; see [41] for other numerical examples previously obtained with the Mumford-Shah regularization.

In Figure 4.6 we show the reconstruction of the noiseless signal of Figure **??** provided information only out of the interval $[100, 150]$ which has to be restored. On the left boxes we show the results due to algorithm (4.37) and on the left ones the solution computed by iterative soft-thresholding. In the former the solution is again piecewise smooth and in the latter a (almost) piecewise constant solution is instead produced.

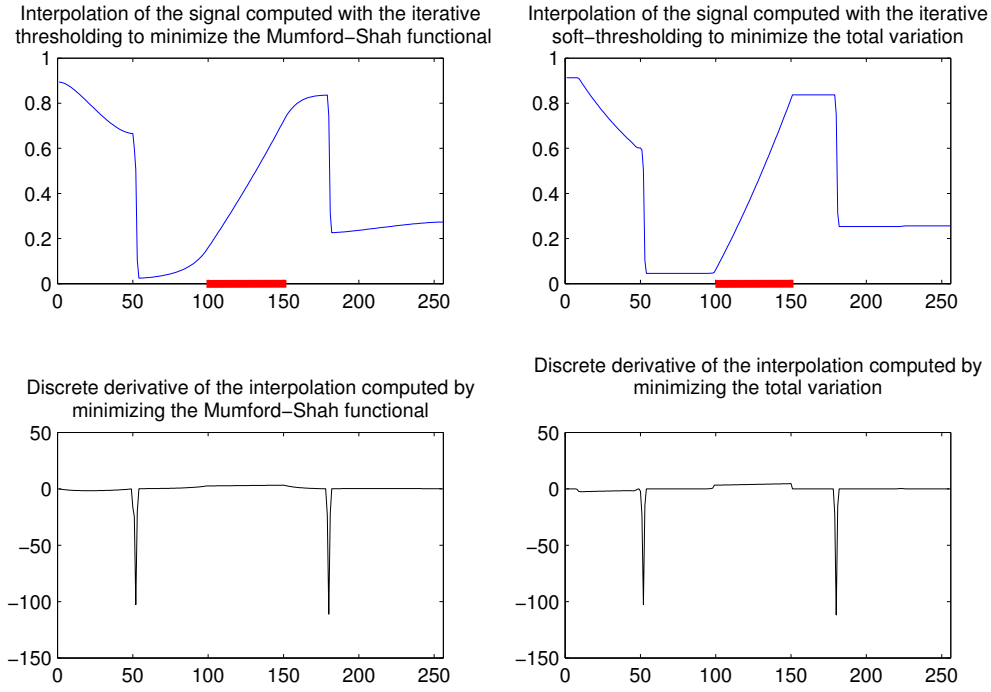Figure 4.6: Interpolation of an incomplete signal by means of the Mumford-Shah regularization and the total variation minimization provided by respective iterative thresholding algorithms. The red interval is the region where no information on the original signal is provided. The thresholding parameters used for the numerics are $r = 2.2$ and $\gamma = 0.002$ for (4.65), and $\gamma = 0.002$ for the soft-thresholding (4.104).

4.8 Appendix

4.8.1 Proof of Proposition 4.2.3

First, we recall Weierstrass' Theorem, which is used in the proof of Proposition 4.2.3 below.

Theorem 4.8.1 (Weierstrass' Theorem). *The set of minima of a convex function f over a subset $X \subset \mathbb{R}^N$ is nonempty and compact if X is closed, f is lower semicontinuous over X, and the function $\tilde{f}$, given by*

$$\tilde{f} = \begin{cases} f(x) & , \text{ if } x \in X, \\ \infty & \text{ otherwise,} \end{cases} \tag{4.66}$$

is coercive, i.e., for every sequence $(x_k) \subset X$ s.t. $\|x_k\| \to \infty$, we have $\lim_{k \to \infty} f(x_k) = \infty$.

The following two lemmas will be helpful in the proof of Proposition 4.2.3.

Lemma 4.8.2. *Let $F(u)$ be a convex function defined on $\mathbb{R}^N$ having the general form $F(u) = \left[u^t A u + b^t u + \sum_{1 \le j \le N} |u_j|^p \right]$, for some $p \ge 1$. Fix x and d in $\mathbb{R}^N$. If F is bounded above and below on the ray $\{x + td, t \ge 0\}$, then F is constant on the line $x + td$.*

Proof. Let $\mu(t) = F(x+td)$, and note that μ is convex because F is convex. Moreover, μ has the general form $\mu(t) = P(t) + \sum_{1 \le j \le N} c_j \|x_j + td_j\|^p$ where $P(t)$ is a polynomial in t of order at most 2. Without loss of generality, suppose $0 \le \mu(t) \le 1$ for all values

182
182

of $t \in \mathbb{R}^+$. Then there exists a sequence of points $(t_n)_{n \in \mathbb{N}}$, $t_n \to \infty$ for $n \to \infty$, for which $\mu(t_n)$ is a convergent sequence; let us denote the limit of this sequence by γ.

1. **Case 1:** $1 \le p \le 2$. To repeat,

$$\lim_{n \to \infty} \mu(t_n) = \lim_{n \to \infty} P(t_n) + \sum_{1 \le j \le N} c_j \|x_j + t_n d_j\|^p = \gamma. \tag{4.67}$$

   Since $0 = \lim_{n \to \infty} \mu(t_n)/t_n^2$, it follows that all coefficients in $\mu(t)$ of degree 2 must vanish. In turn, then, $0 = \lim_{n \to \infty} \mu(t_n)/t_n^p$, has the implication that for each j, one of the coefficients c_j or d_j must vanish as well. Following in the same manner, we conclude that all linear coefficients in $\mu(t)$ also vanish, leaving only the possibility that $\mu(t) \equiv \gamma$ is a constant function.

2. **Case 2:** $p > 2$: The proof in this case is identical to that of the previous case, and as such we leave the details to the reader.

$\square$

Lemma 4.8.3. *Suppose F is a convex function defined on $\mathbb{R}^N$ that is bounded from below, and has the property that if F is bounded above on a ray $\{x + td, t \in \mathbb{R}^+\}$, then F is constant on the line $x + td$. Then if F is constant on the line $x + td$, F is also constant on any parallel line $y + td$.*

Proof. Let $\mu(t) = F(x + td)$ which by assumption is a constant function $\mu(t) = \gamma$, and let $v(t) = F(y + td)$. Fix $t \in \mathbb{R}^+$, and let z be the point $z = x + 2(y - x)$, i.e.

183

$y = \frac{1}{2}x + \frac{1}{2}z$. By convexity of F, we have that

$$F(y + td) = F\left(\frac{1}{2}z + \frac{1}{2}(x + 2td)\right) \leq \frac{1}{2}F(z) + \frac{1}{2}\mu(2t) = \alpha, \qquad (4.68)$$

for a constant α. It follows that F is bounded above by α on the ray $\{y + td, t \in \mathbb{R}^+\}$, from which it follows, by assumption, that F is constant on the line $y + td$. $\qquad \square$

We now prove Proposition 4.2.3. Choosing $x_0 \in X$, we define the (nonempty) set

$$M := X \cap \{x \in \mathbb{R}^N, F(x) \leq F(x_0)\}. \qquad (4.69)$$

Obviously, the set M is convex and closed. By assumption, F is bounded from below on X and hence on M. Therefore, if M is bounded, then Weierstrass' Theorem yields the desired result.

Thus, we may assume that M is unbounded. Then, the convexity of M implies that M contains a ray $r = \{z + td, t \geq 0\}$. Denote by $r_1, r_2, ..., r_J$ a set of J rays in M corresponding to linearly independent vectors $d_1, ..., d_J$, so that any ray in M can be expressed as a linear combination of the $r_1, ..., r_J$. By definition of M and by the assumption, F is bounded on M, hence, F is constant on each of the the lines $z_j + td_j$, according to Lemma (4.8.2). From Lemma (4.8.3), it follows that F is constant along each line $x + td_j$ for arbitrary $x \in \mathbb{R}^N$, from which we deduce that F is constant along any line $x + td$ for arbitrary $d \in Y = \text{span}\{d_1, ..., d_J\}$. Thus, we project X onto the subspace of $\mathbb{R}^N$ that is orthogonal to Y; call this subspace $\tilde{X}$.

From the foregoing arguments, we have

$$\inf_{\tilde{X}} F(u) = \inf_{X} F(u) \tag{4.70}$$

As $\tilde{X}$ is still a convex polyhedral set, and by construction $\tilde{M} = \tilde{X} \cap \{x \in \mathbb{R}^N\}$ contains no rays, Weierstrass' Theorem yields the desired result.

4.8.2 On uniform boundedness of $\|D_h^\dagger\|$

The aim of the second part of the appendix is to prove the uniform bound $\|D_h^\dagger\| \leq 1/2$ eluded to in Section 3.1. Again, $\|A\|$ denotes the spectral norm of the matrix A, and $D_h^\dagger : \mathbb{R}^{n-1} \to \mathbb{R}^n$ is the pseudo-inverse of the discrete derivative matrix D_h as given by (4.6), with the identification $n = \lfloor 1/h \rfloor$. From the expression for D_h, and the knowledge that $D_h D_h^\dagger = I$ is the identity operator and $D_h^\dagger D_h = (D_h^\dagger D_h)^*$ is self-adjoint, the $n \times (n-1)$ matrix $D_h^\dagger$ is identified as follows:

$$D_h^\dagger = \frac{1}{n^2} \begin{pmatrix} -(n-1) & -(n-2) & -(n-3) & \ldots & \ldots & -1 \\ 1 & -(n-2) & -(n-3) & \ldots & \ldots & -1 \\ 1 & 2 & -(n-3) & \ldots & \ldots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & 3 & \ldots & \ldots & n-1 \end{pmatrix}. \tag{4.71}$$

It is well-known that the spectral norm of an $m \times n$ matrix can be bounded by the more manageable entry-wise Frobenius norm, according to

$$\|A\| \le \|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{i,j}|^2}. \tag{4.72}$$

As such, we need only to bound the sum of the squares of the entries of $D_h^\dagger$. The sum $S_n^1 = \sum_{j=1}^{n-1} |d_{1,j}|^2$ over entries in the first row of $D_h^\dagger$ is given by $S_n^1 = (n - 1)(2n - 1)/(6n^3)$, using the familiar formula $\sum_{j=1}^{N} j^2 = \frac{1}{6}N(N + 1)(2N + 1)$. The analogous sum over entries in the jth row of $D_h^\dagger$ is seen inductively to satisfy $S_n^j = S_n^1 - \frac{(j-1)}{n^2} + \frac{j(j-1)}{n^3}$. The total sum $S_n = \sum_{j=1}^{n} S_n^j$ is then $S_n = \frac{1}{6} - \frac{1}{6n^2}$, and we arrive at the desired uniform bound:

$$\|D_h^\dagger\| \le \sqrt{S_n} \le \frac{1}{\sqrt{6}} < 1/2. \tag{4.73}$$

4.8.3 Proof of Proposition 4.3.3

In order to help the reading of the current proof, as well as the proofs of Theorem 4.5.7 and Theorem 4.6.2 in later appendices, we report in Table 1 the notation of the functions used in the proof of Proposition 4.3.3 for the definition of $H_{(p,r)}$.

Consider the functions

$$L_p(t, \lambda) = (t - \lambda)^2 + \min\{|t|^p, r^p\}, \tag{4.74}$$

$L_p(t, \lambda)$	$= (t - \lambda)^2 + \min\{|t|^p, r^p\}$
$G_p(t, \lambda)$	$= (t - \lambda)^2 + |t|^p$
$F_p(t)$	$= t + \frac{p}{2} \operatorname{sgn} t |t|^{p-1}$, $p > 1$
$S_p(\lambda)$	$= G_p(F_p^{-1}(\lambda), \lambda) = (F_p^{-1}(\lambda) - \lambda)^2 + |F_p^{-1}(\lambda)|^p$, $p > 1$
$H_{(p,r)}(\lambda)$	$= \arg\min_{t \geq 0} L_p(t, \lambda)$ for general $\lambda \geq 0$, $p > 1$
	$= \arg\min_{0 \leq t \leq r} G_p(t, \lambda) = F_p^{-1}(\lambda)$ for $0 \leq \lambda \leq r$
	$= \begin{cases} F_p^{-1}(\lambda), & \text{if } G_p(F_p^{-1}(\lambda), \lambda) \leq r^p \\ \lambda, & \text{else} \end{cases}$ for $\lambda > r$.

Table 4.1: Notation of the functions involved in the definition of $H_{(p,r)}$ as in the proof of Proposition 4.3.3.

and

$$G_p(t, \lambda) = (t - \lambda)^2 + |t|^p. \tag{4.75}$$

The proof reduces to solving for

$$H_{(p,r)}(\lambda) = \arg\min_{t \in \mathbb{R}} L_p(t, \lambda) \tag{4.76}$$

as a function of $\lambda \in \mathbb{R}$. Since $L_p(t, \lambda) = L_p(-t, -\lambda)$, the function $H_{(p,r)}(\lambda)$ will be odd, and since also $H_{(p,r)}(0) = 0$, we can, without loss of generality, restrict the domain of interest to $\lambda > 0$. On this domain, $H_{(p,r)}(\lambda) = \arg\min_{t \in \mathbb{R}} L_p(t, \lambda)$ is non-negative, since $L_p(t, \lambda) \leq L_p(-t, \lambda)$ when $t \geq 0$ and $\lambda \geq 0$. Hence, we can restrict the minimization of $L_p(t, \lambda)$ to $t \geq 0$.

It will be convenient to split the proof into two cases: $1 < p$ and $p = 1$.

1. We first analyze the case $1 < p$.

Note that

$$\arg\min_{t\geq r} L_p(t,\lambda) \;=\; \arg\min_{t\geq r}(t-\lambda)^2$$

$$=\; \max\{\lambda,r\}, \tag{4.77}$$

so that the minimization (4.76) naturally splits into the following two cases:

(a) If $\lambda \leq r$, the minimizer has to be searched in $[0,r]$, hence

$$H_{(p,r)}(\lambda) = \arg\min_{0\leq t\leq r} G_p(t,\lambda) = F_p^{-1}(\lambda) \leq \lambda \tag{4.78}$$

where $F_p^{-1}(\lambda)$ is the functional inverse of the increasing, and continuous

function

$$F_p(t) = t + \frac{p}{2}\,\mathrm{sgn}\,t|t|^{p-1}. \tag{4.79}$$

(b) On the other hand, if $\lambda > r$, the minimizer has to be searched in $[0,\lambda]$,

hence

$$H_{(p,r)}(\lambda) \;=\; \begin{cases} F_p^{-1}(\lambda), & \text{if } G_p(F_p^{-1}(\lambda),\lambda) \leq r^p \\ \lambda, & \text{else} \end{cases}.$$

By implicit differentiation of the functional relation $F_p(F_p^{-1}(\lambda)) = \lambda$, it is clear

that the functions $F_p^{-1}(\lambda)$ and $S_p(\lambda) := G_p(F_p^{-1}(\lambda),\lambda)$ are strictly increasing

188

functions in λ. Indeed, we have the bounds

$$0 < \frac{d}{d\lambda} F_p^{-1}(\lambda) = \left(F_p'(F_p^{-1}(\lambda)) \right)^{-1} = \left(1 + \frac{p(p-1)}{2}(F_p^{-1}(\lambda))^{p-2} \right)^{-1} \leq 1,$$

and

$$
\begin{aligned}
\frac{d}{d\lambda} S_p(\lambda) &= \frac{\partial}{\partial t} G_p(F_p^{-1}(\lambda), \lambda) \frac{d}{d\lambda} F_p^{-1}(\lambda) + \frac{\partial}{\partial \lambda} G_p(F_p^{-1}(\lambda), \lambda) \\
&= (2(F_p^{-1}(\lambda) - \lambda) + p(F_p^{-1}(\lambda))^{p-1}) \frac{d}{d\lambda} F_p^{-1}(\lambda) - 2(F_p^{-1}(\lambda) - \lambda) \\
&= 2 \left(1 - \frac{d}{d\lambda} F_p^{-1}(\lambda) \right) (\lambda - F_p^{-1}(\lambda)) + p \frac{d}{d\lambda} F_p^{-1}(\lambda)(F_p^{-1}(\lambda))^{p-1} \geq 0,
\end{aligned}
$$

since $0 \leq \frac{d}{d\lambda} F_p^{-1}(\lambda) \leq 1$, and

$$0 \leq F_p^{-1}(\lambda) \leq \lambda. \tag{4.80}$$

Also observe that $F_p^{-1}(r + \frac{p}{2} r^{p-1}) = r$, and $S_p(r + \frac{p}{2} r^{p-1}) = r^p + \frac{p^2}{4} r^{2p-2} > r^p$. This leads us to immediately conclude that

(i) If $\lambda \leq r$, then $H_{(p,r)}(\lambda) = F_p^{-1}(\lambda)$ (from (4.78)).

(ii) If $\lambda \geq r + \frac{p}{2} r^{p-1}$, then $S_p(\lambda) = G_p(F_p^{-1}(\lambda), \lambda) > r^p$, so that $H_{(p,r)}(\lambda) = \lambda$.

(iii) Since $S_p(r) < r^p$ while $S_p(r + \frac{p}{2} r^{p-1})) > r^p$, the intermediate value theorem implies that there exists a unique value $\lambda'(r,p)$ lying *strictly within* the interval $\left(r, r^{p-1}(\frac{p}{2} + r^{2-p}) \right)$ at which

$$S_p(\lambda') = r^p, \tag{4.81}$$

189

and

$$H_{(p,r)}(\lambda) \;=\; \begin{cases} F_p^{-1}(\lambda) & \lambda < \lambda'(r,p) \\[4pt] \lambda & \lambda > \lambda'(r,p) \end{cases}. \tag{4.82}$$

At λ', $H_{(p,r)}(\lambda') = \arg\min_{t\geq 0} L_p(t,\lambda')$ is not uniquely defined and is realized at $F_p^{-1}(\lambda')$ and at λ'. In this case, we identify $H_{(p,r)}(\lambda') = F_p^{-1}(\lambda)$ for the sequel; as will be made clear, this will not cause problems in the ensuing analysis. Finally, note that

(iv) At λ', the function $H_{(p,r)}$ has a discontinuity $\delta(r,p) = \lambda' - H_{(p,r)}(\lambda')$ that is strictly positive, as long as $r > 0$. Indeed, on the one hand, we know that $\lambda'(r,p) > r$, on the other hand, $H_{(p,r)}(\lambda') < r$. This follows because $H_{(p,r)}(\lambda') = F_p^{-1}(\lambda')$, and

$$(F_p^{-1}(\lambda'))^p < (F_p^{-1}(\lambda') - \lambda')^2 + |F_p^{-1}(\lambda')|^p = S_p(\lambda') = r^p.$$

2. The analysis of the case $p = 1$ is left to the reader since it follows a similar argument as for $p > 1$.

4.8.4 Proof of Theorem 4.5.2

We assume that the operator $T^*T : \ell_2(\mathcal{I}) \to \ell_2(\mathcal{I})$ is nonnegative, so that its spectrum lies within an interval $[\delta, 1]$ with $\delta \geq 0$, and the operator $I - T^*T$ has norm $\|I - T^*T\| \leq 1 - \delta$. In particular, if T^*T is invertible, then the inequality $\delta > 0$ is strict, and so $\|I - T^*T\| \leq 1 - \delta < 1$.

We wish to show that the map $\mathbb{U}_{\mathcal{I}_0}$ with component-wise action

$$[\mathbb{U}_{\mathcal{I}_0} u]_i = \begin{cases} F_p^{-1}([(I - T^*T)u + T^*g]_i), & \text{if } i \in \mathcal{I}_0 \\ ((I - T^*T)u + T^*g)_i, & \text{if } i \in \mathcal{I}_1 \end{cases} \qquad (4.83)$$

is a contraction. To this end, let v, v' be arbitrary vectors in $\ell_2(\mathcal{I})$.

1. If the index $i \in \mathcal{I}_1$, then

$$|[\mathbb{U}_{\mathcal{I}_0}(v)]_i - [\mathbb{U}_{\mathcal{I}_0}(v')]_i| = |[(I - T^*T)(v - v')]_i|;$$

2. If the index $i \in \mathcal{I}_0$, then we split the analysis in two cases $p > 1$ and $p = 1$:

    (a) for $p > 1$, we have

$$\begin{aligned} |[\mathbb{U}_{\mathcal{I}_0}(v)]_i - [\mathbb{U}_{\mathcal{I}_0}(v')]_i| &= \left| F_p^{-1}([(I - T^*T)v + T^*g]_i) - F_p^{-1}([(I - T^*T)v' + T^*g]_i) \right| \\ &= \left| \frac{d}{d\lambda} F_p^{-1}(\xi) [(I - T^*T)(v - v')]_i \right| \\ &< \left| [(I - T^*T)(v - v')]_i \right| \qquad (4.84) \end{aligned}$$

where the second equality is an application of the mean value theorem, which is valid since $F_p^{-1}(\lambda)$ is differentiable. The final inequality above follows from implicit differentiation of the relation

$$F_p^{-1}(F_p(t)) = t$$

and the observation that $|\frac{d}{dt} F_p(t)| > 1$ (see the proof of Proposition 4.3.3);

191

(b) for $p = 1$, by analyzing all cases, we get also that

$$\left|[\mathbb{U}_{\mathcal{I}_0}(v)]_i - [\mathbb{U}_{\mathcal{I}_0}(v')]_i\right| \leq \left|\left[(I - T^*T)(v - v')\right]_i\right| \tag{4.85}$$

Together, we have

$$
\begin{aligned}
\|\mathbb{U}_{\mathcal{I}_0}(v) - \mathbb{U}_{\mathcal{I}_0}(v')\|_{\ell_2(\mathcal{I})}^2 &= \sum_{i \in \mathcal{I}} \left|[\mathbb{U}_{\mathcal{I}_0}(v)]_i - [\mathbb{U}_{\mathcal{I}_0}(v')]_i\right|^2 \\
&\leq \sum_{i \in \mathcal{I}} \left|\left((I - T^*T)v - v'\right)_i\right|^2 \\
&= \|(I - T^*T)v - v'\|_{\ell_2(\mathcal{I})}^2 \\
&\leq \|I - T^*T\|^2 \|v - v'\|_{\ell_2(\mathcal{I})}^2 \\
&\leq (1 - \delta)\|v - v'\|_{\ell_2(\mathcal{I})}^2. \tag{4.86}
\end{aligned}
$$

As $\mathbb{U}_{\mathcal{I}_0}$ is a contraction, we arrive at the stated result by application of the Banach Fixed Point Theorem.

4.8.5   Proof of Lemma 4.5.4

If $i \in \mathcal{I}_1$, then $\bar{u}_i = \bar{u}_i + \left[T^*(g - T\bar{u})\right]_i$, which is satisfied if and only if $\left[T^*(g - T\bar{u})\right]_i = 0$ as stated. It remains to analyze the case $i \in \mathcal{I}_0$, and, again, we split the argument in the cases $p > 1$ and $p = 1$.

1. First suppose $p > 1$. Using the notation $\bar{\lambda} = \bar{u}_i + \left[T^*(g - T\bar{u})\right]_i$, the fixed point characterization (4.51) translates to

$$F_p^{-1}(\bar{\lambda}) = \bar{u}_i.$$

But of course $\lambda = F_p(\bar{u}_i)$ is the unique value at which $F_p^{-1}(\lambda) = \bar{u}_i$, and so this implies that

$$\left[T^*(g - T\bar{u})\right]_i = F_p(\bar{u}_i) - \bar{u}_i, \tag{4.87}$$

and, by reversing operations, the relation (4.87) in turn implies the fixed point condition (4.51).

2. The case $p = 1$, which is similar, is left to the reader.

4.8.6 Proof of Theorem 4.6.1

The proof will be much simplified by the following lemma which characterizes vectors such as $\bar{u}$ that satisfy the fixed point relations (4.61) or (4.62):

Lemma 4.8.4. *If u and v are such that*

$$\mathcal{J}_r^{p,surr}(u + v, u) - \|v\|_{\ell_2(\mathcal{I})}^2 \geq \mathcal{J}_r^{p,surr}(u, u) = \mathcal{J}_r^p(u), \tag{4.88}$$

then $\mathcal{J}_r^p(u + v) \geq \mathcal{J}_r^p(u)$.

Proof. For any u and v, the following holds because $\|L\| \leq 1$:

$$\mathcal{J}_r^p(u + v) = \mathcal{J}_r^{p,surr}(u + v, u) - \|Lv\|_{\ell_2(\mathcal{I})}^2 \geq \mathcal{J}_r^{p,surr}(u + v, u) - \|v\|_{\ell_2(\mathcal{I})}^2. \tag{4.89}$$

If in addition u and v satisfy (4.88), then the desired result is achieved by virtue of the equality $\mathcal{J}_r^{p,surr}(u, u) = \mathcal{J}_r^p(u)$. $\square$

Let us show now the proof of Theorem 4.6.1. By Lemma 4.8.4, it suffices to show

193

that at a fixed point $\bar{u}$ defined by (4.61) or (4.62), any perturbation $\delta h \in \ell_2(\mathcal{I})$ with norm $\|\delta h\|_{\ell_2(\mathcal{I})} \leq \min\{[\lambda'(r,p) - r], [r - H_{(p,r)}(\lambda')]\}$ will satisfy

$$\mathcal{J}_r^{p,surr}(\bar{u} + \delta h, \bar{u}) - \mathcal{J}_r^{p,surr}(\bar{u}, \bar{u}) \geq \|\delta h\|_{\ell_2(\mathcal{I})}^2. \tag{4.90}$$

After expanding the left-hand-side above, the inequality (4.90) is seen to be equivalent to

$$2 \sum_{i \in \mathcal{I}} \delta h_i [T^*(T\bar{u} - g)]_i + \sum_{i \in \mathcal{I}} \left[\min\{|\bar{u}_i + \delta h_i|^p, r^p\} - \min\{|\bar{u}_i|^p, r^p\} \right] \geq 0. \tag{4.91}$$

At this point, it is convenient to consider the summation over $i \in \mathcal{I}_0$ and $i \in \mathcal{I}_1$ separately.

By Lemma 4.5.1, the first summand above vanishes over $\mathcal{I}_1$ and

1. if $1 < p$, then $\sum_{i \in \mathcal{I}} \delta h_i [T^*(T\bar{u} - g)]_i = -\sum_{i \in \mathcal{I}_0} \delta h_i \, \text{sgn} \, u_i \frac{p}{2} |u_i|^{p-1}$;

2. if $p = 1$, then $\sum_{i \in \mathcal{I}} \delta h_i [T^*(T\bar{u}-g)]_i = -1/2 \sum_{i \in \mathcal{I}_0^b} \delta h_i \, \text{sgn} \, u_i + \sum_{i \in \mathcal{I}_0^a} \delta h_i [T^*(T\bar{u} - g)]_i$.

With respect to the second summation, observe from Proposition 4.3.3 that for all $1 \leq p$, $|\bar{u}_i| \geq \lambda'(r,p) > r$ for $i \in \mathcal{I}_1$, so that this summation vanishes over $\mathcal{I}_1$ for any perturbation δh satisfying the component-wise inequality $|\delta h_i| \leq \lambda'(r,p) - r$. Similarly, $|\bar{u}_i| \leq H_{(p,r)}(\lambda') < r$ for $i \in \mathcal{I}_0$, so that for any perturbation δh satisfying component-wise $|\delta h_i| \leq \min\{[\lambda'(r,p) - r], [r - H_{(p,r)}(\lambda')]\}$, we have that

$$\sum_{i \in \mathcal{I}} \left[\min\{|\bar{u}_i + \delta h_i|^p, r^p\} - \min\{|\bar{u}_i|^p, r^p\} \right] = \sum_{i \in \mathcal{I}_0} |\bar{u}_i + \delta h_i|^p - |\bar{u}_i|^p. \tag{4.92}$$

194

The desired result follows if we can show that

1. $1 < p \leq 2$: $\left[|\bar{u}_i + \delta h_i|^p - |\bar{u}_i|^p - \delta h_i p[\text{sgn } u_i]|u_i|^{p-1}\right] \geq 0$, for all $i \in \mathcal{I}_0$

2. $p = 1$:

   (a) $|\delta h_i + \bar{u}_i| - |\bar{u}_i| - \delta h_i[\text{sgn } u_i]] \geq 0$ for all $i \in \mathcal{I}_0^b$, and

   (b) $\delta h_i[T^*(T\bar{u} - g)]_i + |\delta h_i| \geq 0$, for all $i \in \mathcal{I}_0^a$.

The inequality in $2(b)$ follows directly from Lemma 4.5.1; by symmetry, 1 and $2(a)$ follow if, for any $u \geq 0$,

$$\min_{v \in \mathbb{R}} \left[f(v) := |u + v|^p - u^p - pu^{p-1}v\right] = \min_{v \geq -u} (u + v)^p - u^p - pu^{p-1}v \geq 0. \quad (4.93)$$

When $p = 1$, the right-hand-side is identically zero and the result holds. When $1 < p \leq 2$, differentiating the right-hand-side gives that $f(v)$ has a local minimum at $v = 0$, at which $f(0) = 0$, and, at the endpoint, $f(-u) = (p-1)u^{p-1} \geq 0$.

4.8.7 Proof of Theorem 4.6.2

Suppose that u^* is a minimizer of the functional $\mathcal{J}_r^p$. Consider the partition of the index set $\mathcal{I}$ into $\mathcal{I}_0 = \{i \in \mathcal{I} : |u_i^*| \leq r\}$ and $\mathcal{I}_1 = \{i \in \mathcal{I} : |u_i^*| > r\}$, and note that $|\mathcal{I}_1| < \infty$, or else $|\mathcal{J}_r^p(u^*)|$ would not be finite. As in the proof of Theorem (4.5.5), consider the unique decomposition $u^* = u_0^* + u_1^*$ into a vector u_0^* supported on $\mathcal{I}_0$ and another u_1^* supported on $\mathcal{I}_1$. Again, let $\mathcal{P} : u \to u_1$ and $\mathcal{P}^\perp = \mathcal{I} - \mathcal{P} : u \to u_0$ denote the orthogonal projections onto the subspaces $\ell_2^{\mathcal{I}_1}(\mathcal{I})$ and $\ell_2^{\mathcal{I}_0}(\mathcal{I})$, respectively, and consider the operators $T_0 = T\mathcal{P}^\perp$ and $T_1 = T\mathcal{P}$.

By minimality of u^*, if we fix u_0^*, the vector u_1^* satisfies $u_1^* = \arg\min_{z \in \ell_2^{\mathcal{I}_1}(\mathcal{I})} \mathcal{J}_{r,1}^p(z)$, where

$$\mathcal{J}_{r,1}^p(z) := \|T_1 z - (g - T_0 u_0^*)\|_{\ell_2(\mathcal{J})}^2 + \sum_{i \in \mathcal{I}_1} \min\{|z_i|^p, r^p\}. \tag{4.94}$$

Since all coefficients in u_1^* have absolute value $|(u_1^*)_i| > r$, the vector u_1^* also minimizes the functional

$$\|T_1 z - (g - T_0 u_0^*)\|_{\ell_2(\mathcal{J})}^2, \tag{4.95}$$

or, else, the vector z^* minimizing (4.95) would satisfy $\mathcal{J}_{r,1}^p(z^*) < \mathcal{J}_{r,1}^p(u_1^*)$, contradicting the minimality of u_1^*. In fact, u_1^* must be the *unique* vector minimizing (4.95). For, if another vector u' also minimized (4.95), then the operator T_1 would have a nontrivial null space containing the span of some nonzero vector v, so that all vectors in the affine space $\{u_1^* + tv : t \in \mathbb{R}\}$ would be minimal solutions for (4.95). In this case, we would have also the freedom of choosing from this affine subspace a vector u' having one coefficient u_i' satisfying $|u_i'| < r$. But such a vector u' satisfies $\mathcal{J}_{r,1}^p(u') < \mathcal{J}_{r,1}^p(u_1^*)$, contradicting the minimality of u_1^*.

It follows that the operator T_1 must have trivial null space, and u_1^* is the unique minimal least squares solution to (4.95), well-known to be explicitly given by

$$u_1^* = \left(T_1^* T_1\right)^{-1} T_1^* (g - T_0 u_0^*), \tag{4.96}$$

so that $T_1 u_1^*$ is the unique orthogonal projection of $(g - T_0 u_0^*)$ onto the range of T_1.
Actually $\mathcal{P}_1 = T_1 (T_1^* T_1)^{-1} T_1^*$ is the orthogonal projection onto the range of T_1, due
to the non-triviality of the null space of T_1. Therefore we have $T_1 u_1^* = \mathcal{P}_1 (g - T_0 u_0^*)$.
It easily follows that

$$T_1^* \big(T_1 u_1^* - (g - T_0 u_0^*) \big) = 0, \tag{4.97}$$

or, in other words,

$$\big[T^* (g - T u^*) \big]_i = 0, \text{ for all } i \in \mathcal{I}_1. \tag{4.98}$$

Now, on the other hand, by observing that any optimal variable u_1 for fixed u_0
depends on u_0 via the relationship $u_1 = \big(T_1^* T_1 \big)^{-1} T_1^* (g - T_0 u_0)$, we easily infer that
the vector u_0^* minimizes

$$\mathcal{J}_{r,0}^p(v) \;=\; \| \mathcal{P}_1^\perp (T_0 v - g) \|_{\ell_2(\mathcal{J})}^2 + \sum_{i \in \mathcal{I}_0} \min\{ |v_i|^p, r^p \}, \tag{4.99}$$

where $\mathcal{P}_1^\perp$ denotes the orthogonal projection operator onto the orthogonal comple-
ment of the range of T_1.

Consider the convex functional,

$$\mathcal{F}(v) := \| \mathcal{P}_1^\perp (T_0 v - g) \|_{\ell_2(\mathcal{J})}^2 + \| v \|_{\ell_p^{\mathcal{I}_0}(\mathcal{I})}^p, \tag{4.100}$$

and note that $\mathcal{J}_{r,0}^p(u) \le \mathcal{F}(u)$, while at the same time $\mathcal{J}_{r,0}^p(u_0^*) = \mathcal{F}(u_0^*)$ by virtue of
the fact that $|u_i^*| < r$. For $p > 1$ it follows that u_0^* is also a minimizer of $\mathcal{F}(u)$, and

197

so satisfies the Euler-Lagrange equations [7],

$$\left(T_0^* \mathcal{P}_1^{\perp}(T_0 u_0^* - g)\right) + \frac{p}{2} \operatorname{sgn} u_0^* |u_0^*|^{p-1} = 0, \tag{4.101}$$

which imply the fixed point conditions

$$\left[T^*(g - Tu^*)\right]_i = \frac{p}{2} \operatorname{sgn}(u_0^*)_i |(u_0^*)_i|^{p-1}, \text{ for all } i \in \mathcal{I}_0. \tag{4.102}$$

For $p = 1$ one uses results from [30] to conclude that

$$u_0^* = \mathbb{S}_{1/2}(u_0^* + T_0^* \mathcal{P}_1^{\perp}(g - T_0 u_0^*)), \tag{4.103}$$

where $\mathbb{S}_\gamma$ is the so-called *soft-thresholding*, defined component-wise $\mathbb{S}_\gamma(v) = (S_\gamma(v_i))_{i \in \mathcal{I}}$, where

$$S_\gamma(\lambda) = \begin{cases} 0, & |\lambda| \leq \gamma \\ \lambda - \frac{\operatorname{sgn}\lambda}{2}, & |\lambda| > \gamma. \end{cases} \tag{4.104}$$

(Actually, [30, Proposition 3.10] only states that any fixed point of (4.103) is a minimizer of (4.100); nevertheless the converse also holds, see [43, Remarks (1), pag. 2515].) The fixed-point condition (4.103) implies

$$\begin{cases} \left[T^*(g - Tu^*)\right]_i \in [-1/2, 1/2], & |u_i^*| \leq 1/2 \\ \left[T^*(g - Tu^*)\right]_i = 1/2 \operatorname{sgn} u_i^*, & 1/2 < |u_i^*| \leq r. \end{cases} \tag{4.105}$$

It remains to verify that

- $|u_i^*| \geq \lambda'(r, p)$, if $i \in \mathcal{I}_1$, and

- $|u_i^*| \leq F_p^{-1}(\lambda'(r, p))$, for $p > 1$, and $|u_i^*| \leq r - 1/4$, for $p = 1$, if $i \in \mathcal{I}_0$.

We show these conditions for $p > 1$ only, as the case $p = 1$ is proved with an analogous argument.

1. We first show that $|u_i^*| \geq \lambda'(r, p)$ if $i \in \mathcal{I}_1$. From the first part of the proof, we know that at a minimizer u^*, the functional $\mathcal{J}_r^p(u^*)$ can be written as

$$\mathcal{J}_r^p(u^*) = \|\mathcal{P}_1^\perp (T_0 u_0^* - g)\|_{\ell_2(\mathcal{K})}^2 + \|u_0^*\|_{\ell_p^{\mathcal{I}_0}(\mathcal{I})}^p + |\mathcal{I}_1| r^p \qquad (4.106)$$

Note that at this point we make explicit use of the finite cardinality of $\mathcal{I}_1$. Fix $i \in \mathcal{I}_1$ and any perturbation $h = h_i e_i$, $h_i \in \mathbb{R}$, along the coordinate i (here, e_i is the ith vector of the canonical basis). Consider the rank-one operator $t_i = T\mathcal{P}_i$, where we use $\mathcal{P}_i$ to denote the orthogonal projection onto the one-dimensional subspace spanned by e_i. Observe that $\| t_i u \| = |(u)_i| \| t_i \|$. Since t_i is orthogonal to the argument $\mathcal{P}_1^\perp (T_0 u_0^* - g)$ under the ℓ_2 penalty in (4.106), the minimality condition $\mathcal{J}^p(u^*) \leq \mathcal{J}^p(u^* + h)$ can be written as

$$\|\mathcal{P}_1^\perp (T_0 u_0^* - g)\|_{\ell_2(\mathcal{K})}^2 + \|u_0^*\|_{\ell_p^{\mathcal{I}_0}(\mathcal{I})}^p + |\mathcal{I}_1| r^p$$
$$\leq \|\mathcal{P}_1^\perp (T_0 u_0^* - g)\|_{\ell_2(\mathcal{K})}^2 + \|u_0^*\|_{\ell_p^{\mathcal{I}_0}(\mathcal{I})}^p$$
$$+ \|h_i t_i\|_{\ell_2(\mathcal{I})}^2 + \min\{r^p, |u_i^* + h_i|^p\} + r^p(|\mathcal{I}_1| - 1) \qquad (4.107)$$

which is equivalent to the condition that

$$r^p \leq \|h_i t_i\|_{\ell_2(\mathcal{I})}^2 + \min\{r^p, |u_i^* + h_i|^p\} \qquad (4.108)$$

199

hold for all $h_i \in \mathbb{R}$. Now, since $\|T\| < 1$, it follows that $\|t_i\| \leq 1$, and (4.108) implies that

$$r^p \leq h_i^2 + \min\{r^p, |u_i^* + h_i|^p\} \tag{4.109}$$

holds for all $h_i \in \mathbb{R}$, or, after the change of variables $\alpha = u_i^* + h_i$, that

$$r^p \leq (\alpha - u_i^*)^2 + \min\{r^p, |\alpha|^p\} \tag{4.110}$$

holds for all $\alpha \in \mathbb{R}$. In particular, the inequality (4.110) must hold at the value α^* that minimizes the right-hand-side. But we already know from Proposition 4.3.3 that such a minimizer α^* is of the form:

$$\alpha^* = \begin{cases} F_p^{-1}(u_i^*), & |u_i^*| \leq \lambda'(r, p) \\ u_i^*, & |u_i^*| > \lambda'(r, p) \end{cases} \tag{4.111}$$

Now, suppose $|u_i^*| < \lambda'(r, p)$. (We know that $|u_i^*| > r$, so then $r < |u_i^*| < \lambda'(r, p)$). From the proof of Proposition 4.3.3 we know that the function $F_p^{-1}(\lambda)$ is increasing, so then $\alpha^* = F_p^{-1}(u_i^*) < F_p^{-1}(\lambda') < r$. Since also S_p is strictly increasing, it follows that $S_p(\alpha^*) < S_p(F_p^{-1}(\lambda')) \leq S_p(\lambda') = r^p$. In the last inequality we used (4.80). (See also Table 1 for recalling the notations used here.) But this is a contradiction to the minimality condition, (4.110), and so we must conclude that $|u_i^*| \geq \lambda'(r, p)$.

2. We now show that $|u_i^*| \leq F_p^{-1}(\lambda'(r, p))$, if $|u_i^*| \leq r$. Recall that for $i \in \mathcal{I}_0$, the

coefficient u_i^* satisfies the fixed point condition,

$$\left[T^*(g - Tu^*)\right]_i = \frac{p}{2}\operatorname{sgn} u_i^* |u_i^*|^{p-1}. \tag{4.112}$$

Fix $i \in \mathcal{I}_0$, and consider as before any perturbation $h = h_i e_i$ along the coordinate i, $h_i \in \mathbb{R}$. Let t_i be the rank-one operator as defined before. Then, the minimality condition $\mathcal{J}_r^p(u^*) \leq \mathcal{J}_r^p(u^* + h)$ is easily seen to be equivalent to

$$
\begin{aligned}
\|Tu^* - g\|_{\ell_2(\mathcal{K})}^2 + |u_i^*|^p \quad \leq \quad & \|Tu^* - g + h_i t_i\|_{\ell_2(\mathcal{K})}^2 \\
& + \min\{r^p, |u_i^* + h_i|^p\} \\
= \quad & \|Tu^* - g\|_{\ell_2(\mathcal{K})}^2 + \|h_i t_i\|_{\ell_2(\mathcal{I})}^2 + 2h_i \langle t_i, Tu^* - g\rangle \\
& + \min\{r^p, |u_i^* + h_i|^p\} \\
= \quad & \|Tu^* - g\|_{\ell_2(\mathcal{K})}^2 + \|h_i t_i\|_{\ell_2(\mathcal{I})}^2 - 2h_i \frac{p}{2}\operatorname{sgn} u_i^* |u_i^*|^{p-1} \\
& + \min\{r^p, |u_i^* + h_i|^p\} \tag{4.113}
\end{aligned}
$$

and the final equality follows directly from the fixed point condition (4.112). Now the chain of inequalities (4.113) implies the minimality condition

$$
\begin{aligned}
|u_i^*|^p \quad \leq \quad & \|h_i t_i\|_{\ell_2(\mathcal{I})}^2 - 2h_i \frac{p}{2}\operatorname{sgn} u_i^* |u_i^*|^{p-1} + \min\{r^p, |u_i^* + h_i|^p\} \\
\leq \quad & h_i^2 - 2h_i \frac{p}{2}\operatorname{sgn} u_i^* |u_i^*|^{p-1} + \min\{r^p, |u_i^* + h_i|^p\}, \tag{4.114}
\end{aligned}
$$

or, again using the change of variables $\alpha = u_i^* + h_i$, the inequality

$$|u_i^*|^p \leq (\alpha - u_i^*)^2 - 2(\alpha - u_i^*)\frac{p}{2}\operatorname{sgn} u_i^* |u_i^*|^{p-1} + \min\{r^p, |\alpha|^p\}. \qquad (4.115)$$

Again, the inequality (4.115) should hold for all α by the minimality of u^*. Minimizers α^* of the right-hand-side of (4.115) also are minimizers of

$$\left(\alpha - (u_i^* + \frac{p}{2}\operatorname{sgn} u_i^* |u_i^*|^{p-1})\right)^2 + \min\{r^p, |\alpha|^p\}, \qquad (4.116)$$

which we know to have the form

$$\alpha^* = \begin{cases} F_p^{-1}(u_i^* + \frac{p}{2}\operatorname{sgn} u_i^* |u_i^*|^{p-1}), & |u_i^*| + \frac{p}{2}|u_i^*|^{p-1} \leq \lambda'(r,p) \\ u_i^* + \frac{p}{2}\operatorname{sgn} u_i^* |u_i^*|^{p-1}, & |u_i^*| + \frac{p}{2}|u_i^*|^{p-1} > \lambda'(r,p) \end{cases} \qquad (4.117)$$

But $u_i^* + \frac{p}{2}\operatorname{sgn} u_i^* |u_i^*|^{p-1} = F_p(u_i^*)$, so the above reduces to

$$\alpha^* = \begin{cases} u_i^*, & F_p(u_i^*) \leq \lambda'(r,p) \\ F_p(u_i^*), & F_p(u_i^*) > \lambda'(r,p) \end{cases} \qquad (4.118)$$

As before, the proof proceeds by contradiction. Suppose that $F_p(u_i^*) > \lambda'(r,p)$, so that $\alpha^* = F_p(u_i^*) > \lambda'(r,p)$ and $S_p(\alpha^*) > S_p(\lambda') = r^p$. Note that, by recalling $F_p(u_i^*) = u_i^* - \frac{p}{2}\operatorname{sgn}(u_i^*)(u_i^*)^{p-1}$, we have

$$S_p(\alpha^*) = (u_i^* - F_p(u_i^*))^2 + |u_i^*|^p = |u_i^*|^p + \frac{p^2}{4}|u_i^*|^{2p-2}. \qquad (4.119)$$

Plugging α^* into the right-hand-side of (4.115), noting that $\lambda'(r,p) > r$ so that

202

$|\alpha^*| \geq r$, and rearranging, yields the inequality

$$|u_i^*|^p \leq r^p - \frac{p^2}{4}|u_i^*|^{2p-2} \text{ or } S_p(\alpha^*) = |u_i^*|^p + \frac{p^2}{4}|u_i^*|^{2p-2} \leq r^p. \qquad (4.120)$$

But this contradicts the assumption that the expression in (4.119) be larger than r^p.

Bibliography

[1] Boris Alexeev and Rachel Ward, *On the complexity of mumford-shah type regularization, viewed as a relaxed sparsity constraint*, IEE Transactions on Image Processing, Accepted.

[2] L. Ambrosio, *A compactness theorem for a new class of functions of bounded variation*, Bollettino della Unione Matematica Italiana **VII** (1989), no. 4, 857–881.

[3] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free-Discontinuity Problems.*, Oxford Mathematical Monographs. Oxford: Clarendon Press. xviii, 2000.

[4] L. Ambrosio and V.M. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence.*, Commun. Pure Appl. Math. **43** (1990), no. 8, 999–1036.

[5] P.M. Aziz, H.V. Sorensen, and J. Van Der Spiegel, *An overview of sigma-delta converters*, IEEE Signal Processing Magazine **13** (1996), no. 1.

[6] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constructive Approximation (2007).

[7] M. Bazaraa, H. Sherali, and C. Shetty, *Nonlinear Programming: Theory and Algorithms.*, John Wiley and Sons, 1993, Second edition.

[8] G. Bellettini and A. Coscia, *Discrete approximation of a free-discontinuity problem*, Numer. Func. Anal. Optim. **15** (1994), no. 3-4, 201–224.

[9] A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press, 1987.

[10] T. Blumensath and M. Davies, *Iterative thresholding for sparse approximations*, J. Fourier Anal. Appl. **14** (2008), no. 5-6, 629–654.

[11] P. Boufounos, M. Duarte, and R. Baraniuk, *Sparse signal reconstruction from noisy compressive measurements using cross validation*, IEEE Workshop on Statistical Signal Processing (2007), 299–303.

[12] B. Bourdin and A. Chambolle, *Implementation of an adaptive finite-element approximation of the Mumford-Shah functional*, Numer. Math. **85** (2000), no. 4, 609–646.

[13] A. Braides, Γ-*Convergence for Beginners*, Oxford Lecture Series in Mathematics and its Applications, vol. 22, Oxford University Press, Oxford, 2002.

[14] A. Bressan, *Unique solutions for a class of discontinuous differential equations*, Proc. Amer. Math. Soc. **104** (1988), no. 3, 772–778.

[15] E. Candes, *The restricted isometry property and its implications for compressed sensing*, Compte Rendus de l'Academie des Sciences, Paris, Series I **44** (2008), 589–594.

[16] E. Candes, L. Demanet, D. Donoho, and L. Ying, *Fast discrete curvelet transforms*, Multiscale Model. Simul **5** (2005), 861–899.

[17] E.. Candes, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Info. Theory **52** (2006), 489–509.

[18] E. Candes and T. Tao, *Decoding by linear programming*, IEEE Trans. Info. Theory **51** (2005), 4203–4215.

[19] E. J. Cands and T. Tao, *The dantzig selector: statistical estimation when p is much larger than n*, Annals of Statistics **35** (2008), 2392–2404.

[20] A. Chambolle, *Un théorèm de Γ-convergence pour la segmentation des signaux*, C.R. Acad. Sci. Paris Série I **314** (1992), 191–196.

[21] ――――, *Image segmentation by variational methods: Mumford and Shah functional and the discrete approximations.*, SIAM J. Appl. Math. **55** (1995), no. 4, 827–863.

[22] A. Chambolle and G. Dal Maso, *Discrete approximation of the Mumford-Shah functional in dimension two*, M2AN Math. Model. Numer. Anal. **33** (1999), no. 4, 651–672.

[23] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems.*, Numer. Math. **76** (1997), no. 2, 167–188.

[24] T. F. Chan and L. Vese, *A level set algorithm for minimizing the Mumford-Shah functional in image processing*, Proceedings of the IEEE Workshop on Variational and Level Set Methods (VLSM'01), 2001, pp. 161–168.

[25] M .Charikar and A. Sahai, *Dimension reduction in the ℓ_1 norm*, Proceedings of the 43rd Symposium on Foundations of Computer Science (2002), 551–560.

[26] A. Cohen, R. DeVore, and W. Dahmen, *Compressed sensing and best k-term approximation*, J. Amer. Math. Soc. **22** (2009), 211–231.

[27] W. Dai and O. Milenkovic, *Subspace pursuit for compressive sensing: Closing the gap between performance and complexity*, 2008.

[28] G. Dal Maso, *An Introduction to Γ-Convergence.*, Birkhäuser, Boston, 1993.

[29] S. Dasgupta and A. Gupta, *An elementary proof of a theorem of johnson and lindenstrauss*, Random Structures and Algorithms **22** (2003), 60–65.

[30] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint.*, Commun. Pure Appl. Math. **57** (2004), no. 11, 1413–1457.

[31] I. Daubechies and R. DeVore, *Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order*, Ann. of Math **158** (2003), no. 2, 679– 710.

207

[32] I. Daubechies, R. DeVore, M. Fornasier, and C. Sinan Gntrk, *Iteratively re-weighted least squares minimization for sparse recovery*, 2008.

[33] I. Daubechies, R. DeVore, C. Sinan Güntürk, and V. Vaishampayan, *A/d conversion with imperfect quantizers*, **52** (2006), 874– 885.

[34] I. Daubechies and O. Yilmaz, *Robust and practical analog-to-digital conversion with exponential precision*, **52, Issue 8** (2006), 3533– 3545.

[35] Ingrid Daubechies, Sinan Güntürk, Yang Wang, and O. Yilmaz, *The golden ratio encoder*.

[36] E. De Giorgi, *Free-discontinuity problems in calculus of variations*, Frontiers in pure and applied mathematics, a collection of papers dedicated to J.-L. Lions on the occasion of his 60th birthday (R. Dautray, ed.), North Holland, 1991, pp. 55–62.

[37] R. DeVore, B. Jawerth, and B. Lucier, *Image compression through wavelet transform coding*, IEEE Trans. Info. Theory **38** (1992), 719–746.

[38] R. DeVore, G. Petrova, and P. Wojtaszczyk, *Instance-optimality in probability with an ell-1 decoder*, 2008.

[39] Thong Do, Lu Gan, Yi Chen, Nam Nguyen, and Trac Tran, *Fast and efficient dimensionality reduction using structurally random matrices*, 2008.

[40] D. Donoho and M. Elad, *Optimally sparse representation from overcomplete distionaries via l_1 norm minimization*, Proc. Natl. Acad. Sci. USA (2003), 2197–2002.

[41] S. Esedoglu and J. Shen, *Digital image inpainting by the Mumford - Shah - Euler image model*, European J. Appl. Math. **13** (2002), 353–370.

[42] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Mathematics and its Applications (Soviet Series), vol. 18, Kluwer Academic Publishers Group, Dordrecht, 1988, Translated from the Russian.

[43] M. Fornasier, *Domain decomposition methods for linear inverse problems with sparsity constraints*, Inverse Problems **23** (2007), 2505–2526.

[44] M. Fornasier and H. Rauhut, *Iterative thresholding algorithms*, Appl. Comput. Harmon. Anal. **25** (2008), no. 2, 187–208.

[45] M. Fornasier and H. Rauhut, *Recovery algorithms for vector valued data with joint sparsity constraints*, SIAM J. Numer. Anal. **46** (2008), no. 2, 577–613.

[46] P. Frankl and H. Maehara, *The johnson-lindenstrauss lemma and the sphericity of some graphs*, Journal of Combinatorial Theory B **44** (1988), 355–362.

[47] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.*, IEEE Trans. Pattern Anal. Mach. Intell **6** (1984), 721–741.

[48] R. Gregorian and G.C.Temes, *Analog MOS Integrated Circuits for Signal Processing.*, Wiley, New York, 1986.

[49] P. Indyk and R. Motwani, *Approximate nearest neighbors: Towards removing the curse of dimensionality*, Proc. 30th Symposium on Theory of Computing (1998), 604–613.

209

[50] G. James, P. Radchenko, and J. Lv., *Dasso: Connections between the dantzig selector and lasso.*, Journal of the Royal Statistical Society, Series B **71** (2009), 127–142.

[51] J.D.Reiss and M.B.Sandler, *The benefits of multibit chaotic sigma delta modulation*, Chaos **11** (2001).

[52] W. Johnson and J. Lindenstrauss, *Extensions of lipschitz maps into a hilbert space*, Contemp. Math **26** (1984), 189–206.

[53] D. Malioutov, M. Cetin, and A. Willsky, *Homotopy continuation for sparse signal representation*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (2005).

[54] R. March, *Visual reconstruction with discontinuities using variational methods*, Image and Vision Computing **10** (1992), 30–38.

[55] J Matousek, *On variants of the johnson-lindenstrauss lemma*, Random Struct. Algorithms **33** (2008), 142–156.

[56] D. Mumford and K. Shah, *Optimal approximation by piecewise smooth functions and associated variational problems*, Commun. Pure Appl. Math. **42** (1989), 577–684.

[57] D. Needell and J. Tropp, *Cosamp: Iterative signal recovery from incomplete and inaccurate samples*, Appl. Comp. Harmonic Anal. (2008).

[58] D. Needell and R. Vershynin, *Signal recovery from inaccurate and incomplete measurements via regularized orthogonal matching pursuit*, 2007.

210

[59] N.S.Jayant and P. Noll, *Digital coding of waveforms: Principles and applications to speech and video*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[60] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc. **73** (1967), 591–597.

[61] Y. Peres and B. Solomyak, *Absolute continuity of bernoulli convolutions, a simple proof*, Math. Research Letters **3** (1996), 231– 239.

[62] R. Ramlau and W. Ring, *A Mumford-Shah level-set approach for the inversion and segmentation of X-ray tomography data*, Journal of Computational Physics **221** (2007), no. 2, 539–557.

[63] L. Rondi, *A variational approach to the reconstruction of cracks by boundary measurements*, J. Math. Pures Appl. **87** (2007), no. 3, 324–342.

[64] ———, *On the regularization of the inverse conductivity problem with discontinuous conductivities*, Inverse Probl. Imaging **2** (2008), no. 3, 397–409.

[65] ———, *Reconstruction in the inverse crack problem by variational methods*, European J. Appl. Math. **19** (2008), no. 6, 635–660.

[66] M. Rudelson and R. Vershynin, *Sparse reconstruction by convex relaxation: Fourier and gaussian measurements*, Proc. 30th Symposium on Theory of Computing (1998).

[67] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms.*, Physica D **60** (1992), no. 1-4, 259–268.

211

[68] P. Shmerkin and B. Solomyak, *Zeros of $\{-1, 0, 1\}$ power series and connectedness loci for self-affine sets*, Experimental Math. (2006).

[69] N. Sidorov, *Almost every number has a continuum of beta-expansions*, Amer. Math. Monthly **110** (2003), 838– 842.

[70] B. Solomyak, *On the random series $\sum \mp \lambda^n$ (an erdös problem)*, Ann. of Math **142** (1995), 611– 625.

[71] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B **58** (1996), 267–288.

[72] J. Tropp and A. Gilbert, *Signal recovery from random measurements via orthogonal matching pursuit*, IEEE Trans. Info. Theory **53** (2007), 4655–4666.

[73] L. Vese, *A study in the BV space of a denoising-deblurring variational problem.*, Appl. Math. Optim. **44** (2001), 131–161.

[74] R. Ward, *Iterative thresholding meets free-discontinuity problems*, Accepted.

[75] _____, *Quiet sigma delta quantization, and global convergence for a class of asymmetric piecewise affine maps*, submitted.

[76] _____, *Robustness properties of golden ratio encoders and beta encoders*, IEEE Transactions on Information Theory **54** (2008), 4324–2334.

[77] _____, *Compressed sensing with cross validation*, IEEE Transactions on Information Theory **55** (2009), 5773–5782.

[78] P. Wojtaszczyk, *Stability and instance optimality for gaussian measurements in compressed sensing*, 2008.

[79] Ozgur Yilmaz, *Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions*, Constructive Approximation **18** (2002), 599–623.

[80] Sidong Zeng, *Global tile attractor of second order single-bit $\sigma\delta$ modulation*, Ph.D. thesis, The City University of New York, 2008.